UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ESTIMACIÓN DE OCUPACIÓN DE CARRERAS UNIVERSITARIAS PARA UNA UNIVERSIDAD PRIVADA ADSCRITA AL SISTEMA ÚNICO DE ADMISIÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

LUIS FERNANDO MEZA JORQUERA

PROFESOR GUÍA:
LUIS ABURTO LAFOURCADE

MIEMBROS DE LA COMISIÓN:
TODD PEZZUTI LLOYD
ALEJANDRA PUENTE CHANDÍA

SANTIAGO DE CHILE
2016

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: Ingeniero Civil Industrial
POR: Luis Fernando Meza Jorquera
FECHA: 4/03/2016
PROFESOR GUÍA: Luis Aburto Lafourcade

Estimación de ocupación de carreras universitarias para una universidad privada adscrita al sistema único de admisión

El sistema de educación superior chileno se encuentra inserto en un modelo competitivo, donde todas las instituciones están generando acciones para captar un mayor número de postulantes, también atraer una mayor cantidad de alumnos con buenos puntajes y poder completar la mayor cantidad de las vacantes que ofrecen al mercado.

Por este motivo, el objetivo de la presente memoria, es el realizar una estimación de ocupación de las carreras durante el período de matrícula estipulada por el DEMRE con la finalidad de poder ejecutar acciones para, desde el primer día aumentar el nivel de ocupación de esas carreras.

En primer lugar se realiza una estimación de la probabilidad de matrícula de los postulantes a nivel individual, para lo que se realizan tres, dos modelos logit, uno base y uno con interacciones, y un árbol de decisión. Para la realización de estos modelos se ocupa información del DEMRE como de la Universidad en estudio para los últimos dos años del proceso 2014 y 2015. De estos modelos se obtiene que las principales variables que afectan de manera positiva la probabilidad de matrícula es la conversión anterior de la carrera, el puntaje ponderado por sobre el promedio de la carrera, la postulación en primera preferencia a la carrera y la variación entre la beca simulada por el alumno y la que le corresponde según sus características. De los modelos realizados, el que obtuvo los mejores resultados, según los estadísticos de clasificación, fue el modelo logit base, con un Recall de 76,6% y un F-Measure de 74,3%, por lo que este modelo se usó para determinar la conversión de las carreras.

Luego se obtiene la conversión estimada por carrera. En este sentido se obtiene un modelo que tiene un error promedio de conversión del 5,7%, el cual está por debajo del error que comete el modelo que tiene la universidad actualmente.

Finalmente, la memoria concluye con recomendaciones sobre aquellas variables que aumentan la probabilidad de matrícula, como la primera preferencia, la interacción entre el alumno y la universidad y sobre como varía la beca por el tipo de alumno y también por el día de matrícula en que se esté. Como trabajo futuro se propone generar un modelo que incluya el efecto que pueda tener la ley de gratuidad sobre el comportamiento de los postulantes hacia la casa de estudio. El trabajo realizado cumple su objetivo, permite estimar la conversión que tendrán las carreras de la universidad, y además identificar las variables más relevantes en la decisión de los postulantes a la hora de matricularse.

Agradecimientos.

A mis padres, Luz y Fernando, a mis hermanos Rossio e Ignacio, que sin su apoyo y esfuerzo durante todos estos años esto no habría sido posible. Gracias por la preocupación, apoyo y compresión que me han brindado en este largo proceso, por esas palabras alentadoras cuando más lo necesitaba y por darme ánimo, sobre todo al comienzo de mi etapa universitaria. También agradecer a mi polola, Yelsi que me acompañó durante todo el proceso de la memoria y me daba ánimo cuando sentía que no podía o no me resultaban las cosas como yo quería.

También quiero agradecer mis amigos y a todas las personas que conocí en esta etapa, personas que siempre estuvieron ahí para darme una mano cuando lo necesité sin pedir nada a cambio, esto habría sido mucho más difícil sin ustedes. Y agradecer a los profesores de la sección, en particular al profesor guía, que me ayudó cuando yo no tenía claro cómo desarrollar mi memoria y finalmente pude sacarla adelante.

Muchas gracias a todos, nada de esto habría sido posible sin ustedes.

Tabla de contenido

Índice de tablas:

Índice de ilustraciones:

Índice de gráficos:

1 ANTECEDENTES GENERALES

1.1 INTRODUCCIÓN

La oferta de educación superior en Chile está dada por tres tipos de instituciones, estas son: Universidades, Institutos Profesionales y Centros de Formación Técnica [1].

Actualmente hay 60 universidades en el país (sin considerar las que están en proceso de cierre) de las cuales 25 pertenecen al Consejo de Rectores (CRUCH) y 35 no pertenecen [2].

En Chile las universidades puedes dividirse en tres tipos de instituciones [2], estas son:

Tabla 1: Tipos de universidades.

Tipo universidad	Descripción
Estatal	Creada por ley, pertenece al Estado, actualmente son 16.
Particulares con aporte estatal	Son universidades creadas antes de 1980 o que derivan de ellas, actualmente 16.
Privadas	Son todas aquellas creadas después de 1980, a partir de la ley número 18962 de 1990, actualmente son 35.

Fuente: Elaboración propia, Mi Futuro.

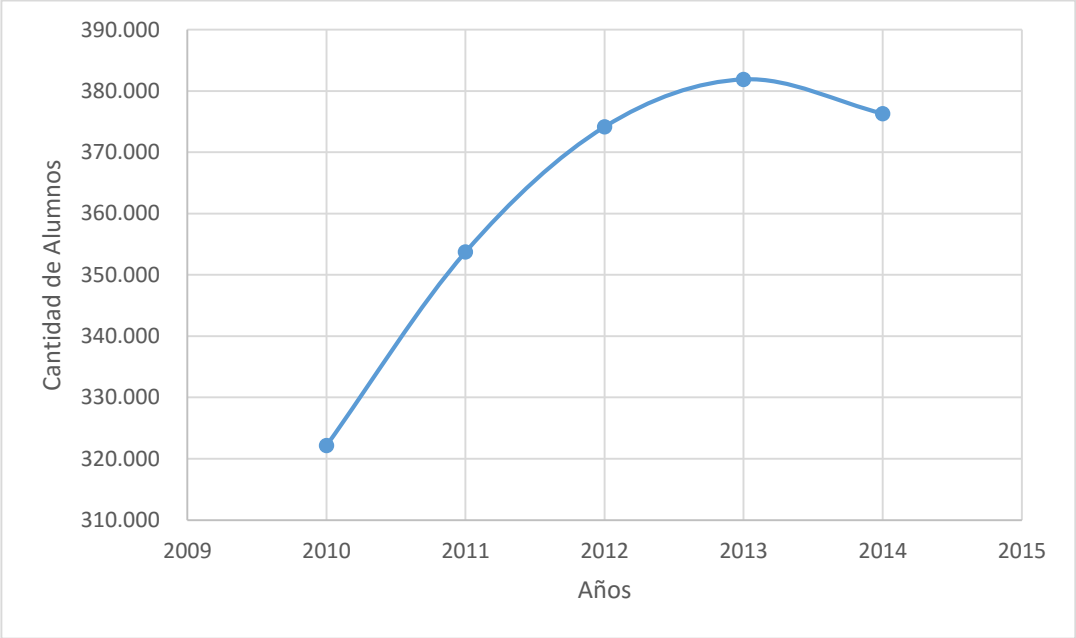
El aumento en las casas de estudio ha estado acompañado también de un incremento en la cantidad de estudiantes que ingresan a la educación superior. En 1986 eran 200 mil los estudiantes que se encontraban cursando educación superior en cualquiera de sus niveles, ya sea universidades, institutos de formación técnica o profesionales [2]. Al año 2014 la cantidad se ha multiplicado llegando a la cifra de 1.215.413 de alumnos presentes en la educación superior, según el Servicio de Información de Educación Superior del MINEDUC (SIES) [3], siendo más de la mitad de ellos, 709.854 alumnos, los que pertenecen a la educación universitaria, tanto pública como privada.

Estos datos muestran el desarrollo y cobertura que ha alcanzado la educación superior en el país, y cada año el número sigue aumentando. Entre los años 2010 y 2014 la cantidad de matriculados en las universidades aumentó un 24,1% [3], y un 16,8% fue para las universidades privadas. Lo que demuestra el gran crecimiento que este tipo de instituciones está alcanzando en el mercado.

De igual forma a medida que aumenta el número de estudiantes que ingresan al sistema de educación superior lo hace también la competencia que deben enfrentar

las instituciones que se encuentran en el mercado, en particular la universidad en estudio, la cual compite tanto con las universidades que pertenecen al CRUCH como con las que no. Esta competencia se ve acrecentada dado el mercado que abarca, ya que cuenta con presencia en tres regiones del país y en cada una de ellas debe hacer frente a la competencia.

Gráfico 1: Evolución matrícula 1° año Universidades Privadas



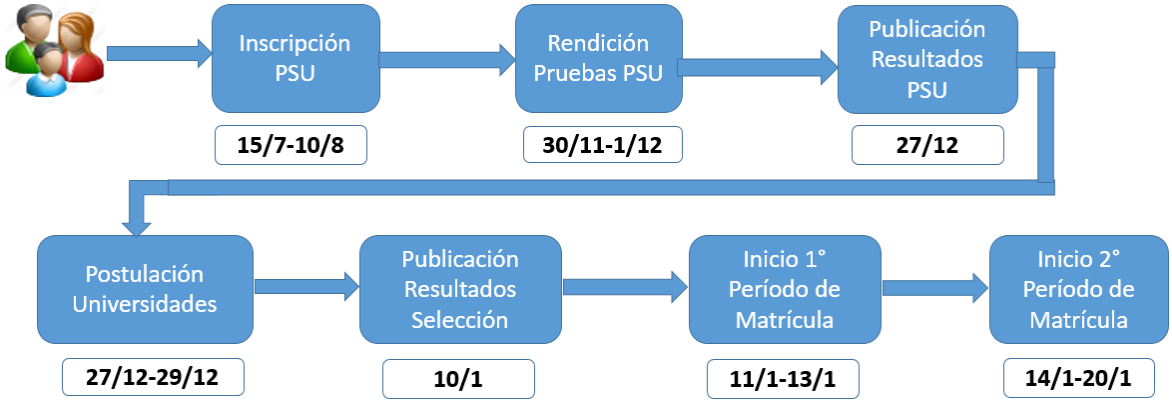
Fuente: Elaboración propia, datos SIES.

Por lo que dada la evolución que ha tenido la educación superior y la fuerte competencia que se enfrenta año a año es cada vez más necesario el crear una marca potente para poder diferenciarse de las otras instituciones, ser más atractivo para el público objetivo con la finalidad de abarcar una mayor parte del mercado. De hecho, la creación de marca es una de las principales estrategias que ocupan las empresas para crear una ventaja competitiva [4].

El sistema actual de selección universitaria contempla dos formas, una donde las universidades no tienen plazos definidos para la matrícula de sus alumnos, y el otro es mediante una prueba de selección universitaria (PSU) a la cual están adscritas 31 universidades [5]. Al pertenecer a este sistema de selección y admisión universitaria las instituciones, para poder llenar sus vacantes a través de la matrícula de estudiantes, se rigen por un mismo proceso, el cual tiene una duración de 10 días desde que se entregan los resultados de la postulación [6].

A continuación se presenta un diagrama del proceso de admisión 2015-2016 desde el momento de la inscripción para rendir la prueba. Donde se muestra que el primer período de matrícula dura tres días, donde se matricula la mayor cantidad de postulantes, ya que el segundo período es para matricular a quienes quedaron en lista de espera durante el primer período.

Ilustración 1: Proceso admisión PSU.



Fuente: Elaboración propia, DEMRE.

Este dato es muy relevante, ya que la universidad en estudio ingresó a este sistema de selección y admisión en el año 2011 [7], por lo que se ha tenido que ir adaptando año a año a este nuevo sistema. Lo cual no ha estado exento de dificultades para la institución, ya que en más de un proceso de admisión ha quedado con vacantes sin poder completar, como en el año 2013 que solo pudo matricular al 64% del total de convocados [8].

Otro aspecto que es relevante conocer es saber el por qué una persona escoge una determinada carrera, qué es lo que la motiva a elegirla y en cuál universidad por sobre la misma carrera en otra casa de estudio u otra carrera, poder identificar cuáles variables inciden en su decisión. Si bien hay muchos aspectos que influyen en el proceso de elección, y alguno de ellos son imposibles de conocer, existen otros factores que si se pueden analizar, y ya se ha hecho en diversos estudios, tanto en Chile como en el extranjero. Dentro de estas variables que pueden ser analizadas podemos diferenciar claramente dos grupos, variables de entorno y variables de marketing.

Las variables de entorno son propias de los postulantes, por ende, imposible de modificar por algún actor del mercado universitario. Sin embargo, las variables de marketing si pueden ser modificadas por las universidades, entre ellas encontramos:

- Beneficios económicos.
- Publicidad sobre la institución.
- Difusión de sus programas a los estudiantes.

Cada una de estas variables de marketing está determinada por una decisión que debe tomar la universidad, por lo cual, tiene un control sobre ellas y puede intervenirlas según lo estime conveniente.

En Chile se han realizado estudios sobre qué variables determinan que una persona escoja la carrera de pedagogía, realizado por Mizala [9], donde encuentran que los

indicadores académicos del alumno resultan ser las variables más significativas para explicar el comportamiento de una persona al elegir una carrera/área por sobre otra, se observa que cuando el puntaje PSU aumenta con respecto al nivel de referencia la probabilidad de que el alumno elija la carrera de pedagogía disminuye drásticamente. También se observa que el colegio de procedencia influye en la elección, ya que personas que provienen de colegios pertenecientes al quintil inferior aumenta la probabilidad, y esta va disminuyendo a medida que el colegio va subiendo su posición en cuanto a quintil. Otras variables que afectan la decisión son factores socioeconómicos del hogar de procedencia como el ingreso y nivel de estudio de los padres. También se encuentra evidencia de que los salarios futuros son un factor importante a la hora de tomar decisiones, esto porque los alumnos ven la educación superior como una inversión, por lo que los retornos que esta traerá son importantes y hacen que este efecto influya en la decisión. Este aspecto no es menor, ya que según la encuesta Casen 2011, una persona con estudios universitarios tiene un salario 2 a 3.5 veces mayor que aquellos que solo tienen enseñanza media [2].

Otro estudio elaborado en Chile realizado por Munita [10], también para analizar qué factores influyen en la probabilidad de que una persona escoja la carrera de pedagogía y se encontró que variables como el género de la persona influye en esta decisión, es más, si es hombre la probabilidad de elegir pedagogía disminuye un 3,3%, y que tendrá mayor probabilidad de elegir carreras como ingeniería o ciencias. En este estudio también se encontró evidencia de que el tipo de colegio afecta, si el alumno proviene de un colegio con GSE (grupo socioeconómico) alto tiende a optar por carreras del área de economía y leyes y no por el área de educación.

A nivel internacional también se han realizado estudios para identificar cuáles factores influyen en la elección de carreras. Se encuentra el trabajo de Chapman [11], el cual encuentra que esta decisión es afectada por características de la persona y por factores externos y estos pueden ser dividido en tres grupos: I) influencia de personas importantes, II) características de la institución, y III) publicidad realizada por la universidad. Se menciona que el aspectos socioeconómico tiene una influencia, el ingreso familiar, nivel de educación de los padres, también influye su desempeño académico, ya que estudiantes con buenos antecedentes académicos son alentados a seguir estudiando.

El estudio realizado por Jiménez y Salas [12], también plantean que el ingreso futuro es un factor que incide en la elección de una determinada carrera, y esto va acompañado del mercado laboral de la carrera medido en la empleabilidad que esta tiene, ya que las personas ven la educación superior como una inversión. Este estudio se realiza en España y se ven las diferencias entre elegir una carrera de 3 años vs 4 años, también menciona que los costos de la educación, directos e indirectos, son una variable relevante. Esto se observa también con el efecto de las becas, según Barrientos [13], encuentra que la conversión, postulantes que se matriculan en la institución, aumenta entre un 20 y 30% cuando se otorga beca, por lo que este también es un factor relevante que influye en la decisión de elección de carrera.

El proyecto puede verse también como una primera aproximación a un modelo de *attribution,* pues al igual que este se busca identificar cuáles son las variables claves que determinan que un postulante se matricule en la casa de estudio y poder así medir su efecto sobre el alumno.

1.2 JUSTIFICACIÓN DEL PROYECTO

Debido a la reciente inclusión de la universidad al sistema único de admisión, donde el período para matricular alumnos es limitado, la universidad en estudio ha tenido dificultades para completar las vacantes ofrecidas en algunas de sus carreras durante el primer período de matrícula.

Este comportamiento se ha evidenciado durante los últimos tres años del proceso, por lo que no es algo fortuito, sino que un patrón que se está repitiendo en el tiempo.

A continuación se muestra la conversión promedio de las vacantes al finalizar el primer período de matrícula para los últimos tres procesos.

Como se observa en la tabla 2 la conversión durante los primeros tres días de matrícula del año 2015 fue menor que la del año 2014, pero para este año, 2014, se tuvo un presupuesto mayor, de un 17% más para otorgar becas a los postulantes en relación al año 2015.

Tabla 2: Conversión primeros tres días de matrícula.

	Año 2015	**Año 2014**	**Año 2013**
Conversión	52,3%	53,3%	48,6%

Fuente: Elaboración propia, datos Penta Analytics.

Como se aprecia la ocupación promedio de las carreras al finalizar el primer período de matrícula es cercano al 50%, por este motivo la universidad debería generar acciones para aumentar la tasa de ocupación de sus carreras.

La universidad a lo largo del año realiza distintas actividades de difusión para posicionarse como marca y que los postulantes la vean como una opción válida a la hora de ingresar a la educación superior. Entre las acciones que se llevan a cabo para lograr este objetivo está por ejemplo actividades en colegios, donde realiza charlas, les presenta a los alumnos los programas que imparte y donde se da la oportunidad de interactuar con ellos resolviendo dudas. Además, la universidad en pos de ayudar a los alumnos, realiza como parte de su plan de difusión, ensayos PSU para que estos puedan practicar y estén más familiarizados con la prueba que tendrán que rendir a fin de año.

Otra herramienta que la universidad utiliza para estar en contacto con los postulantes, es a través del simulador de becas que tiene en su portal web, en el cual los postulantes rellenando unos campos como el puntaje que espera obtener en las

pruebas de la PSU, la carrera a la que postula y el percentil de ingreso en el cual se encuentra le permite conocer el monto de la beca que la universidad le ofrece (sujeto a las condiciones antes declaradas) si decide matricularse en la institución.

Este factor, beca, influye en la decisión de los postulantes, por lo que la universidad otorga esta ayuda económica para aumentar el incentivo de matricularse en la ella. Sin embargo, el atraer un alumno después del primer período de matrícula es más costoso para la universidad, ya que la beca promedio obtenida por estos alumnos es mayor que la de los alumnos que se matriculan durante los primeros tres días. Por ejemplo, para el año 2015 la beca promedio para un alumno que se matriculó después del día tres fue un 15% más alta que para los alumnos que se matricularon durante el primer período. Para el año 2014 la beca promedio para los matriculados durante este período fue un 13% mayor. Junto con esto, cuando la matrícula durante los primeros tres días es alta, el gasto presupuestario extra que realiza la universidad es menor. Para el año 2014 la matrícula durante este período fue de 8200 alumnos aproximadamente por lo que el gasto adicional del presupuesto fue de un 6%, por otra parte, para el año 2015 durante el mismo período la matrícula fue de 7000 alumnos aproximadamente y el gasto adicional fue de un 22%.

Dados los datos antes expuestos la finalidad que tiene el proyecto es de ayudar a la universidad a aumentar la conversión, es decir, que se matriculen más postulantes durante el período de matrícula DEMRE, atrayendo además alumnos con puntajes más altos. Con lo que al aumentar el número de postulantes matriculados en este período ayudaría a disminuir los costos de atraer un alumno fuera de este período, lo que sería un ahorro para la universidad.

Actualmente la empresa Penta Analytics cuenta con un modelo similar, el cual será utilizado como un benchmark para comparar los resultados obtenidos.

¿Ahora por qué es importante realizar el proyecto?, ¿De qué le sirve a la universidad saber que un alumno no se va a matricular? Como se mencionó anteriormente, existen variables de marketing que se pueden manejar, una de ellas es la beca que se ofrece al alumno, por lo que si se sabe que un alumno no se va a matricular se le puede aumentar la beca y, dado que esta es una variable importante en la decisión del postulante, se aumenta el incentivo y ese postulante que no se iba a matricular termine por hacerlo. Junto a esto, el proyecto ayudará a predecir la ocupación de las carreras de la universidad, lo que permitirá realizar un mayor esfuerzo para completar las vacantes en aquellas donde la ocupación se pronostique baja o carreras estratégicas para la universidad, con el fin de aumentar la cantidad de matriculados.

1.3 DESCRIPCIÓN DEL PROYECTO

El proyecto consiste en desarrollar una metodología que le permita a la universidad estimar, a priori, la probabilidad de matrícula de los postulantes a las carreras de la universidad usando diversas fuentes de información, entre las que se encuentran:

- Información socioeconómica de los postulantes
- Información de mercado de las carreras.
- Información sobre los beneficios económicos.
- Información de interacción entre la universidad y los postulantes.
- Información de las preferencias de los postulantes.
- Información histórica de la carrera y universidad.

Además de estimar la probabilidad de matrícula de los postulantes, se obtendrá la conversión de los programas que ofrece la universidad para los tres días de matrícula y también la conversión estimada por cada uno de estos días. Esto se realizará para el primer período de matrícula, en particular el primer día de admisión, para focalizarse en aquellas carreras que tengan una baja matrícula esperada y así poder generar y ejecutar acciones desde el primer día con la finalidad de aumentar su ocupación durante este período.

La metodología para llevar a cabo este proyecto consiste en el desarrollo de un modelo de elección discreta con variables explicativas, las que se obtienen a través de la información declarada por el mismo postulante cuando se inscribe en proceso de admisión PSU. Esta información corresponde al aspecto socioeconómico del alumno y su grupo familiar, a las aptitudes académicas, colegio de procedencia, además de información proporcionada por la universidad a la empresa sponsor sobre la actividad que ha tenido la universidad con el alumno.

Luego estas variables serán agregadas al modelo para predecir la probabilidad de matrícula de cada postulante y los resultados serán comparados con un modelo para el proceso matrícula. Además, se procederá a evaluar este en calidad de ajuste y en su capacidad predictiva para luego generar recomendaciones sobre los postulantes y carreras sobre las que hay que realizar acciones para aumentar su matrícula.

Dentro de las acciones que se pueden realizar se encuentra la priorización de los postulantes para efectuar los llamados a través del call center, y realizarlo en orden descendente según el score que tenga cada uno, también se pueden generar campañas específicas, como que los directores de carrera llamen a los alumnos con buenos puntajes y ordenarlos de acuerdo a su probabilidad de matrícula. Se puede realizar una priorización de las carreras, determinar aquellas que tendrán una baja matrícula y generar campañas focalizadas en esas carreras para aumentar su matrícula. Junto a esto se puede determinar a los alumnos más sensibles a los beneficios económicos y asignarles becas para fomentar su matrícula. Esto es importante ya que se cuenta con un presupuesto acotado, y no se puede entregar becas a todos los postulantes, por lo que la priorización es fundamental.

1.4 OBJETIVOS

1.4.1 OBJETIVO GENERAL

Estimar la probabilidad de matrícula de un postulante convocado por la universidad durante los tres primeros días de matrícula DEMRE.

1.4.2 OBJETIVOS ESPECÍFICOS

Para cumplir con el objetivo general anteriormente planteado es necesario cumplir con una serie de objetivos específicos, los cuales se detallan a continuación.

- Identificar qué variables explicativas se ocuparán en el modelo.
- Generar un modelo para estimar la ocupación de vacantes de las carreras de la universidad.
- Evaluar y validar el modelo y determinar su capacidad predictiva.
- Comparar con modelo anterior desarrollado en la empresa sponsor.
- Entregar recomendaciones para aumentar la probabilidad de matrícula según los resultados del modelo.

1.5 ALCANCES

Como ya se ha mencionado anteriormente la memoria consiste en desarrollar un modelo predictivo de estimación de demanda por carreras universitarias, para lo cual el proyecto tendrá los siguientes alcances:

- La información necesaria para el desarrollo del proyecto será provista por la empresa Penta Analytics, por lo que para el proyecto no se tiene en cuenta el levantamiento de nueva información a través de encuestas, focus group, etc., por lo tanto las variables a utilizar quedan limitadas por la información con la que se cuenta para desarrollar el proyecto.
- Los datos que serán utilizados para la realización de la memoria son los que entrega el DEMRE y que todos los postulantes deben completar para poder rendir la Prueba de Selección Universitaria. Esta información abarca desde los años 2013 hasta el 2015. Además se cuenta con datos a nivel agregado de acceso público para una ventana de 10 años.
- La estimación de demanda es sólo para carreras que la institución ya se encuentra impartiendo en sus facultades, no se considera la estimación de demanda para apertura de nuevas carreras o de carreras que ya impartan en facultades donde no se estén dictando actualmente. Además, la estimación se realizará para la carrera en una sede en particular.
- La memoria no contempla la implementación de la metodología.

- Se estimará la ocupación de vacantes para todas las carreras que se ofrezcan durante el año 2015.

- Se desarrollará un modelo estático, no dinámico. No toma en cuenta la matrícula del día anterior para predecir la del día de hoy.

- Se desarrollará un modelo agregado para los tres días, y el mismo modelo será replicado para generar uno por día, para los primeros tres días de matrícula.

- La memoria no contempla el diseño de acciones para aumentar la matrícula en aquellas carreras que el modelo prediga con baja conversión.

- La memoria no incluye el efecto que podría tener la ley de gratuidad recientemente aprobada

1.6 RESULTADOS ESPERADOS

Los resultados esperados del proyecto son:

- Obtener un conjunto de variables que sean relevantes para el alumno en el momento de su decisión de matricularse o no en una determinada carrera y que puedan ser agregadas al modelo.

- Desarrollar un modelo predictivo que estime la ocupación que tendrán las carreras ofrecidas por la universidad dentro del primer período de matrícula.

- Se espera que el modelo desarrollado tenga un buen nivel de ajuste a los datos y también que su capacidad predictiva sea alta, en pos de equilibrar ambos puntos.

- Es esperable que el modelo desarrollado entregue mejores resultados, con mayor ajuste a los datos y menores errores en su predicción que los desarrollados anteriormente

- Una vez obtenidos los resultados se pretende entregar recomendaciones sobre posibles formar de aumentar la probabilidad de matrícula de los alumnos según los resultados del modelo.

2 MARCO CONCEPTUAL

2.1 MODELO ELECCIÓN DISCRETA

Para la realización del modelo predictivo se utilizará un modelo estructural de elección discreta Logit, ya que éstos permiten la modelización de variables cualitativas [15]. Se utiliza un modelo estructural ya que se propone que la elección de una cierta carrera universitaria es un proceso racional y no probabilístico.

2.1.1 MODELO LOGIT BINOMIAL

Para poder utilizar un modelo de elección discreta las alternativas deben cumplir con las siguientes condiciones [15]: deben ser mutuamente excluyente, es decir deben ser distintas entre sí, también deben ser exhaustivas, todas las posibles alternativas deben estar incluidas dentro de las opciones, y finalmente se debe tener un número finito de alternativas.

Lo que se busca con el modelo Logit [15], es poder capturar la probabilidad P_{ni} de que un agente n (n=1,...,N) elija la alternativa i (i=1,...,I) dentro del set de opciones que se le presentan según los atributos de cada una de ellas. Para lo cual se asume que los agentes son racionales y que ellos buscan maximizar su utilidad con su elección.

Esta utilidad u_{ni} es conocida para el agente, pero no para el analista, por lo que esta se puede descomponer en un factor que es determinístico v_{ni} y un factor que es aleatorio ε_{ni} con lo que la expresión para la utilidad del agente viene dada por:

$$u_{ni} = v_{ni} + \varepsilon_{ni}$$

Con esto la expresión matemática de la probabilidad de que el agente elija una alternativa por sobre otra viene dada de la siguiente forma:

$$P_{ni} = \Pr(u_{ni} > u_{nj}, \forall\, i \neq j)$$
$$P_{ni} = \Pr(v_{ni} + \varepsilon_{ni} > v_{nj} + \varepsilon_{nj}, \forall\, i \neq j)$$

Dentro de la componente determinística se agrupan todas las características observables del agente que pueden afectar o influir en la elección de una alternativa, y la componente aleatoria se pueden modelar como errores aleatorios según una cierta distribución, que para el modelo Logit estos errores se distribuyen según un valor extremo tipo 1, y estos son independientes e idénticamente distribuidos (iid).

Dentro de los tipos de modelos Logit se encuentran el Logit Binomial, en el cual una alternativa se escoge o no, por ejemplo si se compra café o no. Este tipo de decisión puede ser modelo con este tipo de Logit.

Teniendo esto en cuenta y clara la distribución que siguen los componentes aleatorios del modelo Logit, la probabilidad P_{ni} del Logit Binomial se puede obtener de la siguiente expresión.

$$P_{ni} = \frac{e^{v_{ni}}}{1 + e^{v_{ni}}}$$

Entre las propiedades del Logit se encuentran patrones de substitución proporcionales, dentro de ellos está la propiedad de independencia de alternativas irrelevantes (IIA) el cual implica que al agregar una nueva alternativa, similar a una existente en set de opciones, la probabilidad de elegir cada alternativa en el nuevo conjunto es la misma, no hay una elección proporcional de las dos alternativas similares, lo que sería más lógico. Esto es debido a que los errores son idd, lo que se cumple también para las alternativas. También se encuentra la substitución proporcional, es decir, si ante la variación de un atributo en una alternativa, la probabilidad de elección aumenta o disminuye en X%, todas las alternativas aumentarán o disminuirán su probabilidad de elección en X%.

2.2 ÁRBOLES DE DECISÓN

Los árboles de decisión son modelos que permiten predecir una variable dependiente. Existen dos tipos: Árboles de regresión y Árboles de clasificación [16].Se diferencian en que en el primero la variable dependiente en continua, en cambio en el segundo la variable depende es discreta, binaria en esta oportunidad, por lo que para la memoria se utilizará este tipo de árbol.

La construcción del árbol se podría explicar mediante un proceso recursivo, donde se escoge una variable para ubicarla en la raíz del árbol y luego se generan una rama por cada posible valor de la variable [17], y luego re repite el mismo procedimiento para todas las variables.

Los árboles de decisión deben cumplir con dos reglas [18]. La primera de ellas es que debe contar con un solo nodo raíz, es decir, no puede haber más de un nodo al comienzo, y la segunda regla es que cada nodo debe tener un solo nodo padre, a excepción del nodo raíz que no tiene padre.

Para el crecimiento del árbol existen distintos métodos, entre ellos el método de crecimiento CHAID (detección automática de interacciones mediante chi-cuadrado, por sus siglas en inglés), que fue utilizado en la memoria. El este método en cada uno de sus pasos selecciona la variable predictora que tiene una más alta interacción con la variable que se quiere predecir [19], y se utiliza la prueba chi-cuadrado para determinar cuál nodo es el mejor, a mayor valor del estadístico más interacción tiene esa variable con la anterior y se elige ese camino.

2.3 VARIABLES EXPLICATIVAS

Las variables explicativas sirven para ver cómo puede variar el comportamiento de las personas según cierto tipo de variables, como demográficas, cambios en el ambiente, estacionalidades, variables propias de cada individuo como edad, género, educación etc [20].

Estas variables se incluyen dentro de la probabilidad de elección en el modelo Logit de la siguiente forma:

$$P_{ni} = \frac{e^{\beta' X_{ni}}}{\sum_j e^{\beta' X_{nj}}}$$

Donde el vector X_{ni} contiene las variables asociadas al individuo i y que serán utilizadas como variables explicativas de su comportamiento, y el β contiene los coeficientes asociados a las variables, es decir, cuál es el peso de la variable dentro de la elección del agente, puede ser negativo o positivo, si es negativo significa que esa variable impacta negativamente la probabilidad de elegir esa alternativa, y si es positiva quiere decir que la variable afecta positivamente la probabilidad de elección de la alternativa en cuestión.

Con lo cual, luego de estimar los coeficientes β se puede comparar, dadas ciertas características personales, que persona tendrá mayor probabilidad de elegir una alternativa.

El vector de variables explicativas puede ser separado según las características de estas [9]. Se puede generar un vector con la información personal del agente, como por ejemplo edad, género, ingreso, etc. y otro vector que agrupe información sobre su desempeño académico, como puntaje obtenido en la PSU, promedio de notas, etc. Al hacer esta separación la parte determinística de la utilidad queda definida de la siguiente forma:

$$v_{ni} = X_i\beta + Z_i\delta$$

Siendo X_i el vector de características personales y Z_i el que contiene su desempeño académico.

2.4 EVALUACIÓN DE MODELOS

Para la evaluación de modelo se pueden utilizar distintas métricas de bondad de ajuste y comparación de modelos [21], dentro de los que podemos encontrar:

- Criterio de información de Akaike: $AIC = -2Ln(L) + 2K$
- Criterio de información Bayesiano: $BIC = -2Ln(L) + 2Ln(N)$
- Índice de ratios de verosimilitud: $\rho = 1 - \frac{LL(\hat{\beta})}{LL(0)}$

Donde

* L es el estimador de máxima verosimilitud
* K es el número de parámetros
* N es el número de observaciones
* $LL(\hat{\beta})$ es la log-verosimilitud del modelo con todas las variables explicativas
* $LL(0)$ es el modelo restringido, solo estimándose el termino constante

En el caso de los dos primeros criterios se elige el modelo que tenga el menor valor, por lo que se elegirá ese modelo por sobre el otro. En el tercer caso un valor cercano a 0 indica que el modelo con variables explicativas es muy similar al restringido, por lo que estas entregan poca información, por lo tanto se busca un índice lo más cercano a 1 posible.

Junto con estos se utilizará el coeficiente de determinación R, el cual indica que porcentaje de la varianza de Y que es explicada por el modelo.

Además se utilizarán tablas de confusión para el modelo Logit, la que se detalla a continuación.

Al obtener los resultados predichos del modelo estos pueden ser catalogados en clases, positivos y negativos, en este caso valores 1 y 0 respectivamente. Ahora como inputs del modelo se le entregaron los valores reales de positivos y negativos, por lo que al hacer el cruce entre los valores reales y los estimados se obtienen cuatro resultados posibles, estos son Verdadero Positivo (VP) que significa que el valor predicho y el real es positivo, Falso Positivo (FP) es que valor el real es negativo, pero el modelo lo predice positivo, Verdadero Negativo (VN) es cuando el valor real y el predicho es negativo y Falso Negativo (FN) que significa que el valor real es positivo pero el modelo lo predice como negativo, con lo cual se construye la tabla de contingencia.

Tabla 3: Modelo de Tabla de Confusión.

Tabla de Contingencia		Valor real de Y	
		Positivo	Negativo
Valor predicho de Y	Positivo	VP	FP
	Negativo	FN	VN

Fuente: Elaboración propia, [22].

A partir de la cual se definen de los siguientes indicadores, según [22].

❖ Tasa de aciertos:

$$TS = \frac{VP + VN}{VP + FP + FN + VN}$$

- Tasa de error:

$$TE = \frac{FP + FN}{VP + FP + FN + VN}$$

- Recall: esta medida corresponde a los valores positivos correctos versus todos los valores positivos reales.

$$R = \frac{VP}{VP + FN} = \frac{VP}{Positivos}$$

- fp rate: este valor corresponde a los valores negativos clasificados incorrectamente sobre el total de negativos reales.

$$fp\ rate = \frac{FP}{Negativos}$$

- Presición: esta medida corresponde a los valores positivos correctos versus todos los valores clasificados como positivos.

$$P = \frac{VP}{VP + FP}$$

- F-Measure: corresponde a la media armónica entre las medidas Recall (R) y Precisión (P). Indica la medida de precisión que tiene un modelo estadístico, toma valores entre 0 y 1.

$$F = 2 * \frac{R * P}{R + P}$$

Finalmente se grafica la curva ROC en dos dimensiones, la cual mide el trade-off que se produce entre los beneficios de aumentar el número de verdaderos positivos sobre el costo de aumentar los falsos positivos, es decir, que no se produzca un sobre ajuste de los datos. La curva se construye graficando en el eje de las ordenadas la medida *Recall* vs la medida *fp rate* en el eje de las abscisas, por lo que en cada predicción se tiene un valor para estas dos medidas y lo que se busca un valor alto en el eje Y y bajo en el eje X.

3 METODOLOGÍA

La metodología para cumplir con los objetivos planteados se basa en la extracción de información útil a partir de los datos con lo que se cuenta a través de minería de datos para poder resolver el problema de negocios. Esta metodología está compuesta de las siguientes etapas.

3.1 SELECCIÓN DE VARIABLES EXPLICATIVAS PARA INCLUIR EN EL MODELO.

El objetivo de esta parte de la metodología es seleccionar familias de variables que sean relevantes para la decisión del alumno de matricularse o no en una determinada

carrera para incluir en el modelo y ver si al ser incluida en el modelo, tiene poder explicativo o simplemente no es significativa.

Las variables que serán incluidas en el modelo se seleccionarán luego de una revisión bibliográfica de trabajos donde se aborden estos temas. Adicionalmente, se cuenta con el conocimiento experto por parte de la empresa sponsor, ya que se han realizado varios proyectos para la universidad en estudio, por lo que pueden inferir variables que afecten el comportamiento del alumno y que no aparezcan dentro de la literatura consultada.

3.2 LEVANTAMIENTO DE HIPÓTESIS

En esta sección se explicará el por qué estas variables fueron escogidas, es decir, las hipótesis que están detrás de cada uno de los factores seleccionados y explicar cuál es la relación que se presume que tiene la variable con el problema que se busca responder, para después validar o rechazar la hipótesis sobre la variable. Incluso puede darse el caso de que se observe un compartimiento que a priori no se esperaba o no se tenía en cuenta al momento de crear la hipótesis.

3.3 TRATAMIENTO DE DATOS

Esta etapa consiste en la revisión de los datos disponibles para saber con qué datos se cuenta para resolver el problema, como es su codificación, si son categóricos o numéricos, etc. para su posterior selección, limpieza y procesamiento de los mismos, además de transformaciones en los casos que sean necesarios, y donde se creará la base de datos que se ocupará para ejecutar el modelo.

La primera parte de esta etapa consiste en seleccionar los datos que, a priori, pueden ser útiles para el desarrollo de la memoria, ya que se cuenta con gran cantidad de ellos, por lo que se debe buscar en las distintas bases de datos la información relevante al problema que se quiere resolver. Además, es donde se va a consolidar la información y se creará la base de datos. En este caso, como se tienen datos de tres años se debe crear una base para cada año, pero esta debe contener la misma información.

Luego sigue la etapa de limpieza, ésta actividad es fundamental pues es donde se "limpia" la base de datos que será ocupada para ejecutar el modelo, por lo que se tiene que tener un especial cuidado, ya que no puede haber datos incorrectos, ya que esto generaría distorsiones en los resultados del modelo.

Junto con esto se deberá hacer las transformaciones que se estimen necesarias sobre los datos, lo que también incluye la creación de nuevas variables para que toda la información relevante se pueda incluir en el modelo.

3.4 ANÁLISIS EXPLORATORIO

Como una forma de comprender el problema, conocer la situación actual en la que se encuentra la universidad, también se busca conocer los datos y saber cómo estos se relacionan con el tema a resolver. Se realiza un análisis exploratorio de ellos, donde se busca encontrar patrones o relaciones que puedan explicar el problema.

En forma más precisa se realizará un análisis bi-variado de los datos con algunas de las variables explicativas antes seleccionadas y la conversión para comprobar que estas tienen relación con el problema y analizar cuál es el tipo de relación, pues como se mencionó algunos párrafos más arriba cada universidad es distinta, por lo que hay que revisar que las variables seleccionadas aporten información y no solo agregarlas porque la literatura lo diga.

3.5 FORMULACIÓN DE MODELO

En esta etapa se busca determinar qué modelo será utilizado para resolver el problema, explicación de sus resultados y validación del mismo.

3.5.1 DESARROLLO DE MODELOS

Se desarrollaran dos modelos, un modelo a nivel individual y otro a nivel individual por día. El modelo a nivel individual será para estimar la probabilidad de matrícula de los postulantes y analizar qué variables explican en mayor grado si un postulante se matricula o no. El otro de modelo a desarrollar es por día, para tener un mayor detalle y poder hacer seguimiento de la matrícula por día.

De igual forma, el modelo a nivel individual también servirá para obtener la conversión que tendrá la carrera, ya que al estimar la esperanza de las probabilidades de todos los alumnos que postulan a la carrera se podrá calcular la conversión que tendrá la carrera en esperanza.

Para obtener el modelo a nivel individual se realizarán dos modelos, un Logit Binomial y un árbol de decisión, para tener una comparación y quedarse con el que entregue mejores resultados, tanto en ajuste como capacidad predictiva. Estos modelos serán desarrollados con las variables explicativas antes seleccionadas, en diferentes combinaciones de las mismas para obtener el mejor modelo, junto a esto se introducirán interacciones entre las variables, para obtener resultados más específicos, por ejemplo entre beca y percentil de ingreso.

Para realizar el modelo por día de matrícula se tomará la cantidad de matriculados el primer día, luego los matriculados hasta el día dos y finalmente los matriculados hasta el día tres, que es el modelo original.

En resumen se realizarán los siguientes modelos:

- Modelo individual
- Árbol de decisión
- Modelo individual con interacciones.
- Modelo individual por día.

El modelo a nivel individual e individual con interacciones sirve para saber qué variables tienen una mayor influencia en la decisión del alumno de matricularse o no, y además tener un conocimiento sobre cómo afecta a cada uno de los postulantes, para poder medir cómo va a reaccionar un alumno que tiene una baja probabilidad de matrícula si le aumento el monto de la beca, y tener un conocimiento sobre a quién aumentarle la beca y a quien no.

Luego, con el modelo individual se busca estimar la conversión por carrera, para saber si una carrera va a completar sus vacantes o si va a tener una demanda baja. Esto servirá para saber sobre cuáles carreras ejercer acciones para aumentar la cantidad de matriculados, dependiendo de las características de cada una, ya que podría darse el caso de que sea más conveniente aumentar la matrícula de una carrera que tenga una conversión media-alta a una que tenga la conversión muy baja.

Los modelos por día son similares a los descritos anteriormente y se realizan, a nivel individual, para observar cómo varían las variables ingresadas al modelo y obtener información que pueda ser de interés para la institución para poder ejecutar acciones en cuanto a las becas y saber en cuál de los días los postulantes son más sensibles, si es que lo hay, y focalizarse en ese día. Este modelo será desarrollado con el que tenga un menor error promedio en la conversión.

Otro motivo por el cual se desarrolla un modelo por día es para medir la conversión de las carreras, es debido a que el modelo puede pronosticar que una carrera será exitosa y completará gran cantidad de sus vacantes, pero eso no pase al momento de la matrícula y la conversión sea menor a la estimada, y en caso de que es ocurra se puedan realizar acciones para aumentar las vacantes disponibles lo antes posible.

3.5.2 EXPLICACIÓN DE RESULTADOS OBTENIDOS

Una vez que se ha definido el modelo, y se han obtenido sus resultados se procede a explicarlos y medir el grado de ajuste que este tiene con los datos según las medidas de error ya mencionadas en el marco teórico.

Además, se procede a revisar la significancia de las variables explicativas incluidas en el modelo y determinar si, para este modelo, son relevantes o no en la decisión de los alumnos de matricularse en una determinada carrera.

3.5.3 CALIDAD DE AJUSTE DEL MODELO GENERADO

En esta etapa se procede a validar el modelo obtenido a través de la tabla de contingencia y la curva ROC explicadas anteriormente. Con esto se busca medir la calidad del modelo generado.

3.5.4 VALIDACIÓN DEL MODELO

Adicionalmente a la medición sobre la calidad de ajuste del modelo, se desea medir la capacidad predictiva del modelo, esto es de suma importancia, ya que es el objetivo principal de la memoria, por lo que se quiere un modelo confiable con una alta capacidad predictiva.

3.6 MODELO POR DÍA

El modelo por día será desarrollado para tener un punto de comparación en la conversión de la carrera entre un año y otro en un día determinado, ya que el modelo podría sobreestimar o subestimar una cierta carrera, y que al tener una conversión estimada al primer día y contrastar con la matrícula real del día que se está evaluando se puede dar cuenta si se está por sobre o bajo lo esperado y poder generar acciones si es menor a lo esperado o destinar los recursos de esa carrera que va por sobre lo esperado hacía otra carrera que vaya más baja según se estime conveniente.

Este modelo será una réplica del modelo que tenga el menor error promedio de conversión. Este mejor modelo saldrá de la comparación entre ellos.

4 DESARROLLO METODOLÓGICO

4.1 SELECCIÓN DE VARIABLES

La selección de las variables que serán introducidas al modelo en busca de que puedan explicar el comportamiento de los postulantes, se obtienen de conjugar tres aspectos, estos son: 1) la literatura consultada, 2) el conocimiento experto y 3) la información disponible.

Dentro de la literatura consultada, y mencionada anteriormente, acorde al tema se encuentra que las variables socioeconómicas, el rendimiento académico del alumno, el género, la publicidad que realiza la universidad y el financiamiento que se le otorga al alumno son variables que influyen en la decisión del alumno si estudiar o no una determinada carrera, estas categorías se desglosan en las siguientes variables:

- Género
- Ingreso

- Colegio de procedencia
- Estudio de los padres
- Puntaje PSU
- Becas
- Difusión sobre el alumno

A estas variables influyentes según la literatura se agregan las consideradas relevantes por los analistas de la empresa Penta Analytics que trabajan en proyectos relacionados con la universidad en estudio, y también variables consideradas relevantes por el alumno y profesor guía. En síntesis, las variables que se agregan al conjunto anterior son:

- Año de egreso
- Simulaciones y cantidad de simulaciones
- Certificado para validar beca interna y cantidad de certificados por carrera.
- Orden de postulación
- Otras postulaciones del alumno
- Difusión y cantidad de exposiciones a difusión
- Variable binaria por carreras de baja matrícula (20% de las carreras con menor matrícula)
- Variación de la carrera a nivel de univerisdades privadas y mercado
- Valor de la carrera (matrícula, arancel de referencia y real)
- Precio promedio de la matrícula de los Institutos Profesionales (se consideraron tres IP´s para esta variable, estos son: DUOC UC, AIEP, INACAP)
- Conversión anterior de la carrera
- Sede donde postula
- Diferencia entre la beca simulada y la beca real que le corresponde al postulante
- Variación de la beca simulada promedio por carrera entre los años 2015 y 2014.
- Variación en la cantidad de convocados de la carrera entre el año actual y el pasado.
- Nivel de competencia de la universidad.
- Área de cocimiento a la que pertenece la carrera
- Ranking, empleabilidad y salario esperado de la carrera.
- Variación en la cantidad de matriculados a nivel de sede y universidad

El tercer punto clave que se debe cumplir para poder utilizar estas variables es que esta información esté disponible, por lo que al revisar las bases de datos con las que cuenta la empresa se puede acceder a la información antes requerida, así que todas estas variables pueden ser introducidas al modelo.

A continuación se procede a explicar cada una de las variables y como se encuentra codificada la información.

La variable género indica si el postulante es hombre o mujer, una variable dicotómica que tiene valores 1 si es hombre y 2 si es mujer.

La variable ingreso indica el ingreso bruto familiar y viene codificada de 1 a 12, donde 1 es el tramo más bajo y 12 el más alto. En anexos, tabla 52, se puede observar los montos entre los cuales se encuentra cada tramo. Para poder pasar estos 12 tramos a los quintiles de ingreso se tomó el valor promedio entre el valor inferior y el superior de cada tramo y luego se dividió por los integrantes de la familia del alumno.

El colegio de procedencia es el tipo de colegio en el que estudió el alumno según la categorización que realizan las universidades, estos se dividen entre colegios particulares, con el identificador 1, particulares subvencionados, con el 2, y colegios municipales, con el 3.

Estudio de los padres define el nivel educacional de los padres y tiene 13 categorías, y está definida tanto para el padre y la madre en la misma variable, por lo que tiene varias combinaciones posibles entre las 13 categorías del padre con las 13 de la madre.

Puntaje PSU es el puntaje obtenido por el alumno en las pruebas que rindió, que tiene un mínimo de tres pruebas y un máximo de cuatro. Como variable se utilizará un ratio que será el puntaje ponderado del alumno por sobre el puntaje promedio de los convocados que postulan a la misma carrera.

La información sobre las becas se desglosa en tres variables distintas que son, la beca simulada en el portal web de la universidad, ya que a través de esta variable se obtiene la beca promedio ofrecida por la universidad al alumno que la simuló, por lo que se cuenta con la información tanto para los que se matricularon como para los que no se matricularon, siempre y cuando el alumno haya simulado la beca. La segunda variable que se obtiene de la información becas es si el alumno cuenta con beca externa provista por el estado. Esta información también se tiene para los alumnos convocados y no solo para los matriculados.

La última variable que se obtiene de esta información es si el alumno convocado está preseleccionado para el CAE, esta información también se encuentra disponible tanto para los que se matricularon como para los convocados que no se matricularon. Por este motivo no se utiliza la beca real que otorgó la universidad a los alumnos matriculados, ya que se pierde información para los convocados que no lo hicieron.

La información sobre difusión se separa en dos variables. Primero una binaria, que define si el alumno estuvo expuesto a actividades de difusión por parte de la universidad, ya sea charlas informativas, ferias o ensayos tipo PSU. La segunda variable que se crea es la cantidad de veces que el postulante estuvo expuesto a actividades de difusión, ya que el alumno puede verse involucrado en más de una actividad.

El año de egreso se refiere a la cantidad de años que pasó desde que el alumno terminó el colegio hasta que postuló a la universidad y fue convocado, tomando en cuenta que si egresó el mismo año se codifica con un 1 y un 0 en caso contrario, codificándola como variable binaria, y así hasta un máximo de tres años anteriores, desde cuatro años o más se agrupa en una sola variable.

En cuanto a las simulaciones se realiza un desglose similar al realizado para la información de difusión. Se generan dos variables, una que es binaria que considera si el postulante simuló en el portal web la beca aproximada que la universidad le ofrecería si se matricula. Esta variable será si el alumno simuló un beneficio en la misma carrera a la que postuló y por la cual fue convocado, ya que se encontró un número no menor de convocados, aproximadamente un 7%, que simula una cierta carrera, pero que es convocado en otra carrera, por lo que aquellos que simulan una carrera y no son convocados en ella podría ensuciar los datos. La segunda variable que se obtiene es la cantidad de veces que un postulante simuló, ya que se considera que a mayor cantidad de simulaciones más interés demuestra en postular a la universidad.

Otra variable que se considera y que tiene relación con la variable simulación antes descrita, es si el postulante que simuló un beneficio en el portal web de la universidad emitió un certificado que valida el beneficio a la hora de matricularse y lo puede hacer efectivo. Además, se obtiene la cantidad de certificados emitidos por carrera.

En cuanto a la información sobre difusión se pueden obtener dos variables. Una es binaria, que indica si el alumno estuvo expuesto a algún tipo de difusión, como ferias, charlas, visitas a la universidad y el pre-unab. La segunda variable que se puede obtener es la cantidad de veces que un postulante estuvo expuesto a alguna actividad.

Una variable que depende de cada carrera a la que se ven enfrentado los postulantes es el valor de esta y que tiene influencia sobre la decisión del postulante, según la bibliografía consultada. Esta consta de tres partes: el costo de la matrícula, el arancel de referencia y el arancel real. Información que es entregada por la universidad.

De igual forma, el orden de postulación indica en qué lugar de preferencia el alumno postuló a la universidad, esta variable está codificada con valores de 1 a 10. Sin embargo, para el año 2015 solo podían postular dentro de las primeras seis preferencias a la universidad en estudio. Lo que se ha reducido a través de tiempo, para el año 2013 se podía postular dentro de la 10 preferencias posibles, luego se redujo para el año 2014 a las primeras ocho preferencias hasta llegar a las seis primeras para el año 2015.

La variable con información de otras postulaciones es una binaria que dice si el alumno quedó en lista de espera en otra universidad a la postuló en un lugar de preferencia superior del que seleccionó a la universidad en estudio. Además se va a obtener información sobre la institución en la cual el alumno se encuentra en lista de espera, es decir, si esta universidad pertenece al CRUCH o es una universidad privada no perteneciente a este grupo, pero si adscrita al sistema único de admisión.

Se creará una variable que agrupe el 20% de las carreras que tienen la menor conversión para el año 2015, será una variable dummy que indica si la carrera pertenece a este grupo o no. Teniendo un total de 127 carreras, el 20% son aproximadamente 25 carreras.

Se agregará información de la evolución de la carrera en el mercado, es decir, la variación que tuvo la carrera para en el último año, por ejemplo, si se quiere predecir el año 2015 se debe conocer la variación que tuvo la carrera en el año 2014 con relación al año 2013, para saber si la carrera viene en alza, es una carrera estable o es una carrera que está en declive.

Otra variable de mercado que se consideró para el análisis fue obtener el precio promedio de la matrícula de los Institutos Profesionales (IP´s). Los IP's que fueron considerados para obtener este precio de referencia fue: INACAP, DUOC UC y AIEP. Esta variable se hará interactuar con las carreras con baja matrícula. Otra variable que se creó toma en cuenta el valor promedio del arancel de los mismos Institutos profesionales.

Se crearon las mismas variables que la descrita anteriormente, pero con información de las universidades privadas y de todas las universidades del sistema universitario que ofrecen la carrera

La conversión anterior es similar a la variación de mercado, pero lo que se busca es saber una historia de la carrera, tener un detalle de cómo le fue a la carrera el año anterior en la universidad.

Con la variable sede donde postula, se desea saber si la sede tiene algún efecto en la decisión del alumno. Además, si el alumno postula a una sede que se encuentra en la misma comuna o no, lo que está midiendo si existe un efecto de distancia a la hora de matricularse.

Otra variable que se considera es la diferencia entre la beca simulada por el alumno en el portal web y la beca real, esta diferencia es debido a que el simulador ofrece una beca promedio según los puntajes estimados por el alumno antes de conocer el puntaje real y la beca real que le correspondería dado los puntajes que obtuvo en la prueba.

En cuanto a la variación en la beca promedio simulada por carrera se toma en cuenta la beca promedio simulada en la carrera entre un año y otro. Esta variable se obtiene de la beca simulada por el alumno y se promedia entre todos los convocados que tuvieron beneficio en su simulación. De igual forma la variación en la cantidad de convocados mide la diferencia que tuvo la carrera en la cantidad de convocados entre un año y el anterior.

También se mide el nivel de competencia mide cuantas universidades ofrecen el mismo programa que la institución en estudio. No toma en cuenta la cantidad de sedes en las que se imparte el programa. Esta variable se obtuvo a nivel de región,

suponiendo que la universidad en sus sedes de Santiago compite solo con universidades que imparten la carrera en esta región. Lo mismo para las sedes de Viña Del Mar y Concepción.

Además, las carreras son caracterizadas según el área de conocimiento a la cual pertenece, por lo que cada una se asigna a un campo determinado, para realizar esta separación de carreras se utilizó la misma que ocupa el Consejo Nacional de Educación (CNED).

4.2 LEVANTAMIENTO DE HIPÓTESIS

Las variables mencionadas en el apartado anterior fueron escogidas ya que se piensa que pueden influir en que un alumno se matricule o no en la universidad. Esto tiene como fundamento la bibliografía consultada y el conocimiento experto generado de trabajar en proyectos con la misma casa de estudio. Estas hipótesis serán ingresadas en los modelos logit, y se validarán o rechazarán según el nivel de significancia, p-valor, que presenten en él. Significancia mayor al 90% serán consideradas para la realización del modelo.

Más específicamente se busca explicar cuál es el motivo de la elección de estas variables y cómo se cree que afecta a la decisión final.

- Género: se utiliza el género del postulante ya que se ha encontrado evidencia, en trabajos anteriores, de que hay carreras en las que existe un mayor número de hombres o de mujeres, por lo que, si una carrera que ha tenido una mayor matrícula de mujeres y esta vez hay más hombres puede ser que finalmente estos no se terminen matriculando en la universidad.

- Ingreso: esta variable es importante ya que las carreras no tienen el mismo valor del arancel, incluso hay ciertas carreras en las que se tiende a matricular alumnos de un ingreso más alto o bajo, depende por ejemplo de la duración de la carrera o de la cotización de esta por los alumnos.

- Colegio de procedencia: se ha encontrado evidencia de que el colegio de procedencia tiene cierta influencia en la decisión del alumno, esto puede deberse que el colegio promueve más cierto tipo de carreras o áreas específicas, o debido a su condición de universidad privada sea más atractiva para cierto grupo de alumno que provienen de colegios de ingresos más altos.

- Estudio de los padres: esta variable es importante, ya que en la literatura se ha encontrado evidencia de que si el padre tiene estudios superiores motiva al hijo a que tenga un nivel de estudios superior al que tiene él o la madre tiene, por lo que se espera que si el padre tiene estudios universitarios va a promover que el hijo/a estudie una carrera universitaria.

- Puntaje PSU: esta variables es para poder obtener el rendimiento del alumno en la prueba, ya que a mayor puntaje PSU tiene más opciones de universidades para postular, por lo que si obtiene un puntaje alto puede ser menos probable que postule a la universidad.

- Beca simulada: se estima que a mayor monto de la beca ofrecida por la universidad al alumno, este tiene más incentivo a matricularse.

- Beca externa: al igual que la beca simulada se estima que si el postulante cuenta con ella tenga más probabilidades de matricularse. Sin embargo, como la beca externa puede usarse dentro de todas las universidades reconocidas por el estado también tiene más opciones para elegir. A pesar de esto, se considera el primer efecto, si tiene este beneficio tiene una mayor probabilidad de matricularse en la universidad en estudio.

- CAE: si el alumno posee ayuda para el financiamiento de sus estudios tiene mayor probabilidad de matricularse en la universidad, tiene la misma desventaja que la beca externa, con CAE el alumno puede matricularse en cualquier universidad reconocida por el estado.

- Difusión sobre el alumno: se cree que si el alumno estuvo expuesto a difusión tiene mayor probabilidad de matricularse en la universidad sobre un alumno que no tuvo información de la universidad.
- Cantidad de actividades de difusión: al igual que la variable difusión se supone que a mayor exposición en cuanto a cantidad de actividades de difusión tiene una mayor probabilidad de matricularse, es decir, un alumno que tuvo una actividad es menos probable que se matricula si es que tuvo dos o más actividades.

- Año de egreso: se supone que a medida que lleva más tiempo de egresado es menos probable de que se matricula, es decir, que quien egresó el mismo año del proceso o anterior, es más probable que se matricule que alguien que terminó cuarto medio hace cinco años.

- Simulaciones: se cree que si un alumno simula los beneficios que podría tener a través del portal web que dispone la universidad, tiene más probabilidades de matricularse en la universidad que alguien que no simuló, ya que al simular se demuestra un interés por la universidad.

- Cantidad de simulaciones: al igual que la hipótesis anterior, se piensa que a mayor cantidad de simulaciones demuestra un mayor interés por la universidad, por lo que alguien que simula cinco veces tiene una mayor probabilidad de matricularse que alguien que simuló una vez.

- Certificados: la hipótesis que se tiene bajo esta variable es que si un postulante genera el certificado para hacer válido el beneficio simulado tiene una mayor propensión a matricularse, ya que si cuenta con él le deben validar la beca que

le fue ofrecida, siempre que mantenga las condiciones de simulación, por lo que tiene el beneficio asegurado con el certificado.

- Cantidad de certificados por carrera: si una carrera tiene muchos certificados quiere decir que hay interés por parte de los postulantes en ella, lo que podría indicar que van a convertir un mayor número de ellos.

- Orden de postulación: el supuesto bajo esta variable es que si un alumno postuló en su primera preferencia a la universidad es más probable que se matricule en la universidad que alguien que postuló en una preferencia más abajo, de igual forma si alguien postuló a la universidad en segunda opción tiene mayor probabilidad de matricularse de alguien que postuló tercero o cuarto a la universidad.

- Otras postulaciones del alumno (lista de espera): similar al orden de postulación, si un alumno postuló a otra universidad en una preferencia superior a la universidad en estudio y además, quedó en lista de espera en esa universidad tiene menor probabilidad de matricularse que un alumno que fue rechazado en su preferencia anterior. Además, se obtendrá información sobre la universidad en la cual quedó en lista de espera en la preferencia anterior, para saber qué tipo de universidad es y si es competencia directa o no.
- Carrera con baja matrícula: se cree que las carreras con baja matrícula se repiten en el tiempo, y que son carreras que no son muy atractivas para los postulantes, por lo que a mayor valor de la matrícula versus la matrícula de los IP's tienen menos probabilidad de matricularse, ya que prefieren otra alternativa, como una carrera técnica profesional.

  Variación de la carrera en el mercado: la hipótesis que está detrás de esta variables es saber si esta es una carrera cotizada, estable o en decrecimiento, ya que si una carrera viene en alza es más probable que la carrera se complete sus cupos con mayor facilidad, en cambio, si una carrera viene a la baja es porque los postulantes la están demandando menos y es menos probable que complete sus cupos. Se tiene esta variable para la variación tanto para las universidades privadas como para el mercado completo.
- Valor de la carrera: la hipótesis detrás de esta variables es que a mayor costo de la carrera menor probabilidad de matrícula.

- Conversión anterior: lo que se supone es que si la carrera tuvo una conversión alta el año pasado en la universidad este año debería tener una ocupación similar, es darle un peso a la historia de la carrera.

- Sede: se puede dar el caso de que una sede sea más cotizada que otra y la conversión de la carrera esté relacionada con este hecho. Además, se agrega la variable de si el alumno postula a una sede que está en su misma comuna, pues se puede dar el caso de que postula a una sede que quede lejos de su casa y en ese caso decida por matricularse en otra universidad.

- Diferencia beca simulada y beca real: se cree que a una mayor diferencia positiva (negativa) entre la beca que el postulante simuló en el portal y la beca real que le correspondería, es decir, que el beneficio sea mayor al que él simuló, la probabilidad de matricularse es más alta (baja) ya que el beneficio es mayor al que él tenía considerado.

- Variación beca simulada: la hipótesis detrás de esta variable es que si la carrera tiene una variación negativa, quiere decir que la universidad disminuyó el presupuesto para la carrera, por lo que entregará menos beca a los postulantes a la carrera, por lo que al tener una beca menor en relación al año pasado la probabilidad de matrícula disminuye, y viceversa, si la variación es positiva, quiere decir que la universidad está destinando más fondos para la carrera y los postulantes podrían acceder a una beca mayor, lo que incentivaría la matrícula.

- Variación de convocados: se supone que el nivel de convocados afecta en la conversión final del programa de forma que si se convoca una mayor cantidad de alumnos, entre un año y otro, en una cierta carrera la conversión va a aumentar, de igual forma, si la convocatoria a la carrera es menor que el año pasado, la conversión debería disminuir.

- Nivel de competencia: bajo esta variable existen dos hipótesis. En primer lugar se cree que si una carrera tiene una mayor competencia la probabilidad de que se matricule es menor, ya que tienen más opciones para elegir. Por otro lado, a un mayor nivel de competencia puede deberse a que la carrera es más atractiva para los postulantes, por eso hay tantas universidades que ofrecen el programa, por lo que aumentaría la probabilidad de matrícula en esa carrera.

- Área de conocimiento: pueden haber ciertas carreras que sean más atractivas para los postulantes y que tengan una mejor conversión por sobre otras carreras, dependiendo del área a la que pertenezca, y al agruparlas se puede identificar este efecto.

Tabla 4 : Resumen levantamiento de hipótesis (1).

Variable	Hipótesis
Género	Poder afectar en distinta forma en las ciertas carreras
Ingreso	A mayor ingreso mayor probabilidad de matrícula
Colegio procedencia	El tipo de colegio puede incidir en la probabilidad matrícula
Estudio del padre	A mayor estudio del padre mayor probabilidad de matrícula
Puntaje PSU	A mayor puntaje aumenta la probabilidad de matrícula
Beca	A mayor beneficio mayor probabilidad de matrícula

Fuente: Elaboración propia.

Tabla 5: Resumen levantamiento de hipótesis (2).

Variable	Hipótesis
Simulación	A mayor simulaciones aumenta la probabilidad de matrícula
Difusión	A mayor difusión aumenta la probabilidad de matrícula
Certificado	Si emite el certificado en su simulación tiene mayor propensión de matrícula
Cantidad certificados por carrera	A mayor cantidad de certificados emitidos que tenga una carrera mayor probabilidad de conversión
Año de egreso	A menor tiempo de egreso mayor probabilidad de matrícula
Orden de postulación	Primeras preferencias aumenta la probabilidad de matrícula
Otras postulaciones	Disminuye la probabilidad de matrícula
Variación mercado	A mayor crecimiento de la carrera en el mercado mayor probabilidad de matrícula
Carrera con baja conversión y valor matrícula IP's	A mayor diferencia entre el valor de la matrícula de la universidad y de los IP's la probabilidad disminuye
Variación Universidades Privadas	A mayor crecimiento de la carrera en las universidades privadas mayor probabilidad de matrícula
Conversión anterior	A mayor conversión anterior mayor probabilidad de matrícula
Valor carrera	A mayor costo de la carrera (matrícula, arancel de referencia y real) menor probabilidad de matrícula
Comuna sede	Postular a una sede de la misma comuna aumenta la probabilidad de matrícula
Diferencia beca simulada y beca real	A mayor diferencia positiva mayor probabilidad de matrícula
Variación beca simulada	A mayor diferencia positiva mayor probabilidad de matrícula
Variación convocados	A mayor aumento en la cantidad de convocados mayor conversión de la carrera.
Competencia	Se tienen dos hipótesis contrarias, una es que aumenta la probabilidad de la matrícula y la otra es que disminuye la probabilidad de matrícula.
Área de Conocimiento	La probabilidad de matrícula cambia dependiendo del área de conocimiento a la que pertenezca

Fuente: Elaboración propia.

4.3 TRATAMIENTO DE DATOS

Dentro del tratamiento de datos se encuentran las acciones de selección, limpieza y transformación de los datos (en caso de ser necesario).

Los datos para la realización del proyecto se basan en los últimos tres años del proceso y se encuentran en diversas bases de datos que posee la empresa y desde dos fuentes principales. Por una parte son los datos oficiales del DEMRE que es quien se encarga del proceso de admisión universitario y datos propios de la universidad que es la empresa sponsor quien los provee.

Entre los datos que se encuentran disponibles para la memoria están los datos que aporta el alumno al momento de inscribirse para dar la PSU y que son datos de características socioeconómicas, que incluye el RUT del postulante, medio por el cual es identificado, además del ingreso familiar, quién conforma su grupo familiar, el colegio del que proviene, región en la que vive, esta es la base de todos los inscritos para rendir la PSU.

Otra de las bases con la que se trabaja es la de postulaciones a las universidades, que contiene toda la información del alumno en cuanto al proceso de postulación, es decir, esta base muestra a que carrera, universidad y preferencia en la que postuló el alumno, cuál fue el puntaje que ponderó en la carrera, si quedó seleccionado, en lista de espera o rechazado por la universidad que eligió y además en qué posición quedó.

Estas dos bases son provistas por el DEMRE, es la institución que se encarga de consolidar toda esta información para luego hacérsela llegar a todas las instituciones que forman parte del proceso. Ambas bases cuentan con un registro aproximado de 250.000 filas por año y por base.

Además se cuenta con información sobre los beneficios externos con que cuentan los alumnos inscritos en el proceso, estos son becas y crédito con aval del estado (CAE), la cantidad de datos son aproximadamente 80000 en la base de las becas y 120000 los registros para quienes quedaron preseleccionados con el CAE.

Luego se tienen las bases de datos que genera la universidad, entre estas se encuentra la lista con los convocados, es decir, con aquellos postulantes que quedaron seleccionados en la universidad a través del proceso de admisión, en la cual el alumno se encuentra identificado por el RUT y cuenta con información sobre la carrera en la que quedó seleccionado, el puntaje ponderado y el lugar en el que quedó, además se tiene el lugar en el que postuló a la universidad y los puntajes que obtuvo en cada una de las pruebas que rindió. Esta base cuenta con 14000 registros por año aproximadamente.

Adicional a esta, se cuenta con la base de los matriculados que contiene a todos los que se matricularon cada uno de los años en cuestión, esta contiene tanto alumnos de primer año como para los de años anteriores, por lo que cuenta con aproximadamente 40000 registros.

La universidad además genera base de datos de los alumnos sobre los que se realizó actividades de difusión, identificados con el RUT del alumno y el colegio. También lleva un registro de los postulantes que tuvieron interacción con la universidad a través del simulador de becas en su plataforma web, identificados con su RUT, la carrera en la cual simuló y el monto ofrecido al alumno. Ambas bases cuentan con alrededor de 240000 registros aproximadamente, ya que las personas pueden simular cuantas veces quiera de igual forma algunos alumnos están expuestos a más de una actividad de difusión, por lo que no son registros únicos.

El siguiente paso es la consolidación de los datos, agrupar todos los que se consideren útiles para el desarrollo del proyecto sacados de las bases de datos antes mencionadas para generar una sola base por año que contenga la variable que se pretende estudiar con las respectivas variables explicativas.

Para generar la base se ocuparán los convocados y matriculados, es decir, cada base tendrá alrededor de 14000 registros únicos. Se debe crear una base por año, y para generar el modelo se deben ocupar las mismas variables por año.

Luego de agrupar los datos se debe realizar una limpieza de los mismos, ya que al cruzar las bases de datos para extraer la información que será utilizada esto genera muchos datos vacíos o nulos, por lo que hay que tener especial cuidado de que no se vaya a pasar ninguno de estos en la base, ya que el modelo podría entregar resultados alterados.

De igual forma, existen datos que están fuera de rango, sobre todo los que completan los mismos alumnos inscritos para la PSU, ya que muchos de estos se completan mediante alternativas, y si el alumno no responde una pregunta automáticamente se autocompleta con un cero y este queda fuera del rango predeterminado, por lo que el modelo podría interpretar este valor como una opción dentro del set de alternativas, lo cual no es válido.

El último paso dentro del tratamiento de datos es la transformación de algunos de ellos. Esto se realiza por ejemplo cuando la diferencia entre los valores de las variables es muy grande, como en los montos de las becas, donde hay quienes tienen un monto de $0 y otro montos de $4.000.000 por lo que esto podría provocar ciertos problemas en el modelo y por eso se aplica una transformación logarítmica que disminuye estas diferencia. Un problema que surge con esta transformación es que para el valor $0 el logaritmo no está definido, por lo que hay que volver a realizar una transformación y pasar de $0 a $1 con lo que esta función queda definida. Otra transformación que se debe hacer es con respecto al ingreso, ya que la información se encuentra en tramos de ingreso bruto familiar de 1 a 12, esta información es codificada por el DEMRE, y a la universidad le es más útil si los alumnos están categorizados por deciles, con lo cual hay que buscar una forma de transformar estos 12 tramos en solo cinco quintiles. Esto se realiza tomando el valor promedio del tramo de ingreso y luego dividiendo por los integrantes del grupo familiar y se obtiene un ingreso per cápita que se clasifica según el quintil correspondiente.

4.4 ANÁLISIS DESCRIPTIVO

Como se mencionó en la metodología se realiza un análisis descriptivo de los datos en primer lugar para comprender el problema que se desea resolver, además para saber cómo se presentan los datos. Junto con esto se pretende encontrar la relación que existe entre las variables explicativas que se van a introducir al modelo y la matrícula de los convocados.

Para partir el análisis se muestra la cantidad de matriculados y los convocados que ha tenido la carrera durante los últimos tres períodos de matrículas, con lo que se obtiene la conversión promedio de la universidad. Ésta se calcula como el promedio de la conversión de las carreras, y se obtiene entre el cociente entre los matriculados en el programa respectivo sobre los postulantes que fueron convocados a ella.

Tabla 6 : Conversión primer período de matrícula.

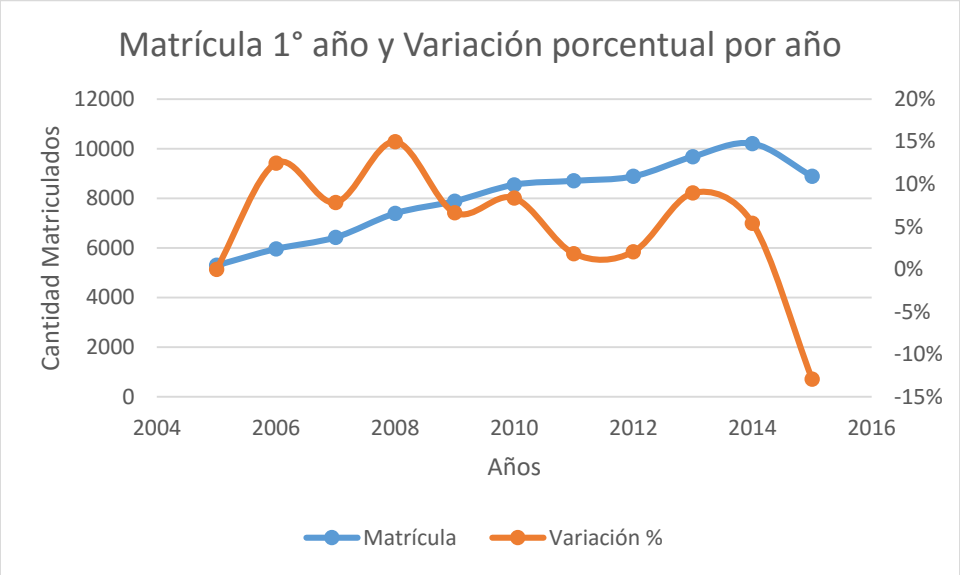
	Año 2015	**Año 2014**	**Año 2013**
Matriculados	7022	8211	7220
Convocados	13216	14806	13906
Conversión	52,3%	53,3%	48,6%

Fuente: Elaboración propia, datos Penta Analytics.

En la tabla 6 se puede apreciar como la ocupación promedio de la universidad ronda el 50%, además se observa que la cantidad de matriculados por años es similar, salvo en 2014 donde tanto la matrícula como los convocados fue mayor, de igual forma, la conversión, donde alcanzó el 53,3%, similar al año 2015, por lo que este es un patrón que se repite a lo largo del tiempo y es estable sin tener una mayor variación a pesar de que se aumente o baje la cantidad de convocados.

A continuación, en el gráfico 2 se presenta como ha sido la evolución de la matrícula de primer año para la universidad en cuestión.

Gráfico 2: Matrícula 1° año a nivel universidad.



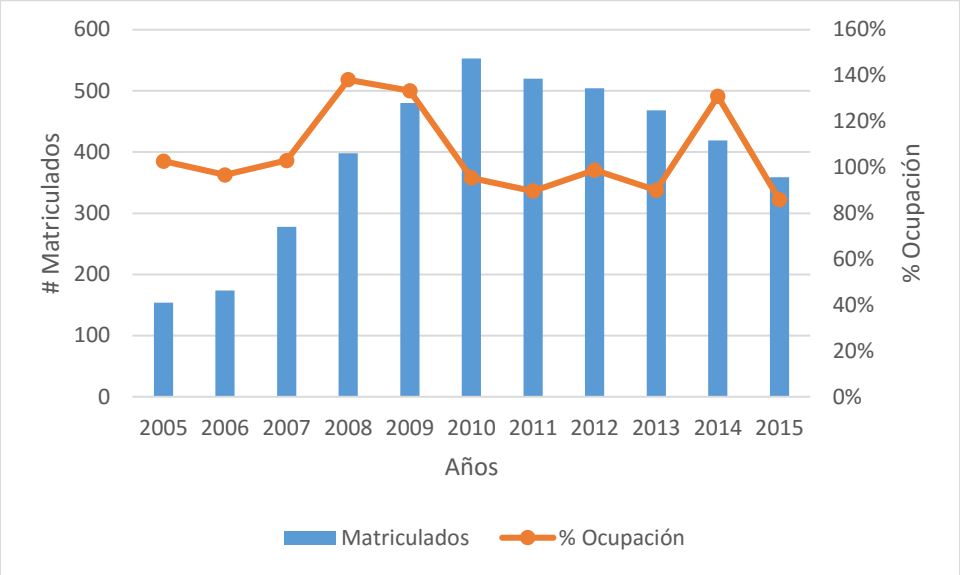
Fuente: Elaboración propia, datos CNED.

En el gráfico se puede observar que la universidad ha tenido una tendencia creciente durante los últimos 9 años, a pesar de que no se tiene un crecimiento estable en el tiempo siempre había tenido un crecimiento a excepción del último año donde se presenta una caída del 13% aproximadamente.

Este descenso, el del último año, se debe a un factor propio de la universidad, ya que la matrícula a nivel del mercado universitario, todas las universidades del sistema, presenta un incremento del 1.2%, por lo que no es un efecto generalizado en las otras universidades.

En cuanto a la evolución por programas se observan comportamientos distintos entre cada uno de ellos. Algunos programas vienen en decrecimiento desde algunos años, otros bajaron el último año siguiendo la tendencia de la universidad y otros aumentaron su matrícula el último año. A modo de ejemplificar estos comportamientos se presentan tres programas en los cuales se puede observar cada una de estas tendencias.

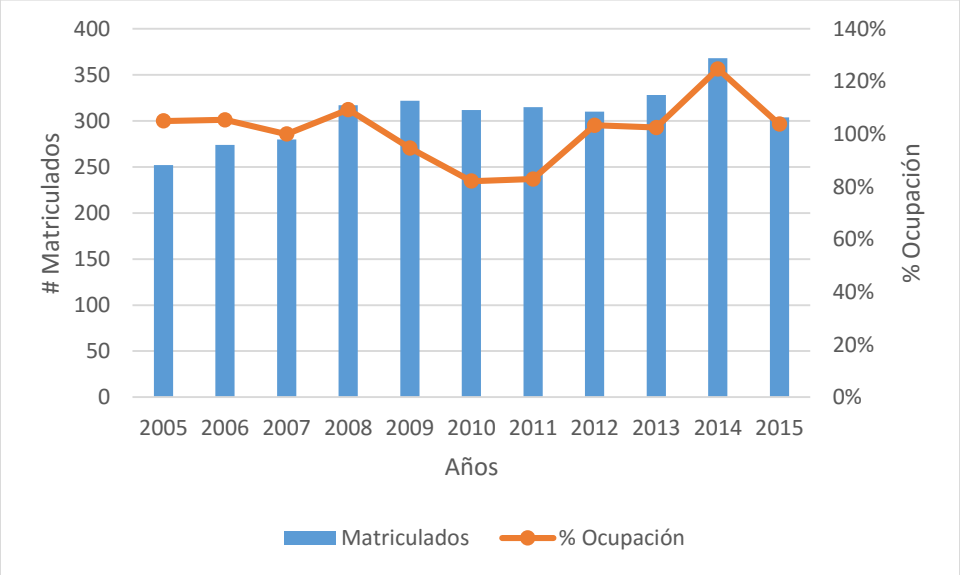
Gráfico 3: Evolución Ingeniería Comercial Casona.



Fuente: Elaboración propia, CNED.

En el gráfico 3 se pueden apreciar la evolución de la carrera de Ingeniería Comercial en la sede Casona, donde se observa que desde el año 2010 la matrícula viene en descenso, de igual forma, el nivel de ocupación para estos años, salvo el 2014, se ha mantenido bajo el 100%, es decir, que las vacantes ofertadas por la universidad para la carrera no se han completado.

Gráfico 4: Evolución Derecho Bellavista.



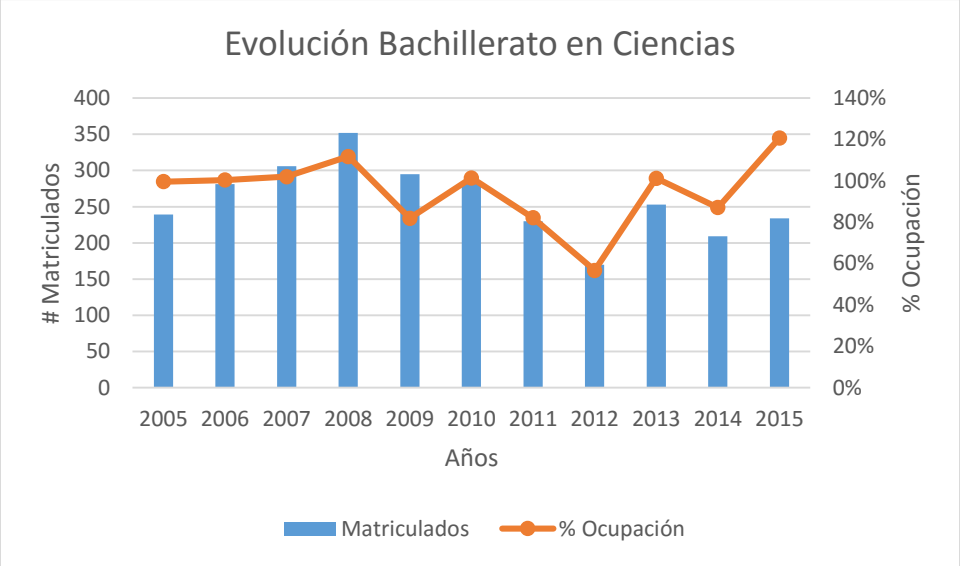
Fuente: Elaboración propia, CNED.

En el gráfico 4 se observa la evolución de la carrera de Derecho en la sede de Bellavista, se puede observar que esta venía en alza desde el año 2012 y que el último año sufrió una baja en la matrícula de 60 alumnos aproximadamente, siguiendo la

tendencia de la universidad este último año. Sin embargo, el nivel de ocupación se ha mantenido alrededor del 100% los últimos cuatro años, a excepción del 2014 donde fue del 120% (gracias a los sobre cupos).

Finalmente se tiene la carrera de Bachillerato en Ciencia en la sede República, que tuvo un aumento en la matrícula para el año 2015, contrario a la tendencia de la universidad.

Gráfico 5: Evolución Bachillerato en Ciencias República.



Fuente: Elaboración propia, CNED.

En el gráfico 5 se observa que esta carrera tuvo un aumento en la matrícula el último año, a pesar de que es una carrera con mucha variación año a año, de igual forma el porcentaje de ocupación está por sobre el 100%.

A continuación se presenta los matriculados y convocados por tipo de colegio de procedencia, cabe destacar que para el sistema universitario estos se clasifican en tres tipos; 1) Particulares, 2) Particular Subvencionado, 3) Municipal.

En la tabla 7 se observa que el tipo de colegio del cual proviene la mayor cantidad de matriculados es de colegios particulares subvencionados, y es un patrón que se repite a lo largo del tiempo, en cuanto a las otros tipos de colegios la cantidad de matriculados es similar para ambos, no se observa una variación muy grande a lo entre un año y otro, salvo para el proceso de admisión 2014 donde la cantidad de matriculados fue mayor.

Adicionalmente, se muestran los matriculados y convocados por tipo de colegio, con lo que se obtiene la conversión por tipo de colegio.

Tabla 7: Conversión por tipo de colegio.

	Año 2015			Año 2014			Año 2013		
Tipo Colegio	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión
Particular	1.346	2.079	65%	1.449	2.188	66%	1.419	2.058	69%
Subvencionado	4.273	8.095	53%	5.045	9.030	56%	4.348	8.432	52%
Municipal	1.403	3.042	46%	1.717	3.588	48%	1.453	3.417	43%
Total	**7.022**	**13.216**	**52,3%**	**8.211**	**14.806**	**53,3%**	**7.220**	**13.906**	**48,6%**

Fuente: Elaboración propia, datos Penta Analytics.

En la tabla se observa que la mayor cantidad de matriculados por tipo de colegio proviene de los particulares subvencionados, lo que se correlaciona con la cantidad de convocados, ya que la mayoría de estos son de este tipo de colegio. Sin embargo, la mayor conversión se da en los colegios particulares, sobre el 65% para los tres años en estudio.

En segundo lugar están los provenientes de colegios subvencionados con una conversión por sobre al 50%y la conversión más baja se da en los colegios municipales, por debajo del 50%, siendo que se convocan en mayor cantidad que a alumnos de colegios particulares y la cantidad de matriculados de ambos tipos de colegio es similar. Por lo que se observa un patrón por tipo de colegio, ya que se presenta una mayor conversión si es un alumno proveniente de uno particular, por lo que el tipo de colegio de procedencia puede influir en la probabilidad de matrícula de un postulante.

Continuando en la dimensión socioeconómica del alumno, la cantidad y distribución de los matriculados según el tramo de ingreso en el cual se encuentran.

En la tabla 8 se puede apreciar que los tramos 2-3 son los que concentran la mayor cantidad de matriculados durante los tres años, a pesar de que para el año 2015 se observa una disminución en el tramo 2. Luego el tramo 4 junto con el 12 son los que concentran la segunda mayor cantidad de matriculados y se observa que el año 2015 la cantidad de matriculados en este tramo es muy similar al año 2014. La mayor cantidad de los matriculados se concentra en estos cuatros tramos de ingreso, de los cuales los 2-3-4 son alumnos que tienen un ingreso bajo a diferencia del tramo 12 que es el más alto, por lo que se presentan dos grupos socioeconómicos muy distintos. Adicionalmente, se observa que el grupo 12 es el que presenta la mayor conversión de los cuatro, por lo que a un mayor ingreso familiar la probabilidad de matrícula aumenta, ya que se aprecia que la conversión aumenta a medida que lo hace el tramo de ingreso. En anexos, tabla 53, se muestran los valores entre los que se encuentra cada uno de los tramos.

Tabla 8: Conversión por tramo de ingreso.

	Año 2015			Año 2014			Año 2013		
Tramo Ingreso	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión
1	361	753	48%	492	1026	48%	507	1206	42%
2	1.405	2980	47%	1.770	3660	48%	1.669	3866	43%
3	1.279	2637	49%	1.518	2922	52%	1.284	2662	48%
4	907	1695	54%	987	1816	54%	836	1532	55%
5	633	1178	54%	748	1274	59%	627	1146	55%
6	404	732	55%	474	768	62%	414	671	62%
7	402	683	59%	514	784	66%	425	674	63%
8	247	386	64%	259	395	66%	221	343	64%
9	109	202	54%	160	249	64%	131	196	67%
10	138	211	65%	137	213	64%	114	178	64%
11	156	232	67%	161	241	67%	129	189	68%
12	981	1527	64%	991	1458	68%	863	1244	69%
Total	**7.022**	**13.216**	**52,3%**	**8.211**	**14.806**	**53,3%**	**7.220**	**13.906**	**48,6%**

Fuente: Elaboración propia, datos Penta Analytics.

En la tabla 9 se presenta la matrícula y conversión por año de egreso.

Tabla 9: Conversión por año de egreso.

	Año 2015			**Año 2014**			**Año 2013**		
Año Egreso	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión
Año proceso	4.065	7.728	53%	4.923	8.899	55%	4.405	8.517	52%
1 año antes	1.917	3.374	57%	2.267	3.790	60%	1.882	3.392	55%
2 años antes	418	812	51%	403	771	52%	371	736	50%
3 años antes	227	427	53%	210	420	50%	188	370	51%
% Matricula	94%	93%		95%	94%		95%	94%	

Fuente: Elaboración propia, datos Penta Analytics.

De la tabla 9 se observa en primer lugar que aproximadamente el 95% de los convocados y matriculados en la universidad tiene una antigüedad de egreso de máximo 3 años, es decir, para el año 2015 el 94% de los matriculados había salido del colegio entre el año 2012 y el año 2015, ocurre lo mismo con los convocados, y este patrón se repite para los tres años en estudio.

Otro aspecto interesante es que la conversión es más alta si el postulante ya tiene un año de egreso que los postulantes que egresaron del colegio el mismo año del proceso y esta también es mayor para los que egresaron hace dos y tres años atrás.

La conversión para todos los años de egreso que se detallan en la tabla es de al menos un 50%. Y en promedio si el alumno egreso el mismo año tiene una conversión mayor que si egresó dos o tres años antes, salvo en 2015 donde la conversión entre los postulantes que egresaron el mismo año del proceso y tres años antes que es la misma.

A continuación se presenta el análisis de la variable "*orden de postulación*" esta variable indica en qué lugar de preferencia el alumno postuló a la universidad, ya que cuenta con 10 opciones para postular.

Tabla 10: Conversión por orden de postulación

	Año 2015			Año 2014			Año 2013		
Orden Postulación	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión
1	4.485	6.687	67%	5.780	8.411	69%	4.717	7.161	66%
2	1.143	2.616	44%	1.181	2.674	44%	1.140	2.693	42%
3	713	1.828	39%	752	1.870	40%	763	1.892	40%
4	342	984	35%	232	810	29%	273	849	32%
5	208	658	32%	136	469	29%	147	547	27%
6	131	443	30%	66	286	23%	87	324	27%
7				31	168	18%	45	200	23%
8				33	118	28%	25	125	20%
9							19	75	25%
10							4	41	10%
Total	**7.022**	**13.216**	**52,3%**	**8.211**	**14.806**	**53,3%**	**7.220**	**13.906**	**48,6%**

Fuente: Elaboración propia, datos Penta Analytics.

De la tabla 10 se observa que para el año 2015 el máximo lugar de preferencia de los matriculados y convocados es el sexto, para el año 2014 el octavo y para el año 2013 hasta el décimo, esto se debe a una política que ha adoptado la universidad estos últimos dos años de no aceptar postulaciones que estén más allá del sexto lugar para el año 2015 y del octavo lugar para el año 2014. Esto tiene relación con que para el año 2015 la cantidad de matriculados en los tramos 4-5-6 haya aumentado.

La conversión para los alumnos que postulan en la primera preferencia es del 67%, 69% y 66% para los años 2015, 2014 y 2013 respectivamente, por lo que se tiene una alta tasa de matrícula de quienes colocan a la universidad como su primera opción.

Además, se nota una gran diferencia entre la conversión de quienes postulan en primera preferencia con quienes postulan en la segunda preferencia, ya que la conversión de estos últimos es 44% para el año 2015. La diferencia entre una preferencia y otra es de al menos 23% para los tres años en estudio.

Y se puede observar que la conversión tiende a decrecer a medida que aumenta el lugar de preferencia en que el postulante selecciona la universidad, llegando a una conversión del 30% para la sexta preferencia en 2015 y a un 10% para la décima preferencia en el año 2013, por lo que postular en primera preferencia aumenta la probabilidad de matricularse en la universidad.

Se aprecia que para estas carreras los convocados el año 2015 fue menor, pero esta disminución es mayor que la cantidad de alumnos que postularon en preferencias más bajas el año anterior, por lo que no se podría decir que la disminución en los convocados se debe a esta restricción, aunque en algunas carreras sumando la

cantidad de postulantes en estas preferencias a los convocados del 2015 se obtiene un número similar a los convocados del año 2014.

Continuando con el análisis de las variables de entrada, se muestra la conversión de los alumnos por las actividades de difusión que realiza la universidad.

Tabla 11: Conversión por difusión.

	Año 2015	Año 2014	Año 2013
Expuesto difusión sobre matriculados	43%	48%	57%
Expuesto difusión sobre convocados	23%	27%	29%

Fuente: Elaboración propia, datos Penta Analytics.

En la tabla 11 se observa que del total de alumnos matriculados solo el 43% estuvo expuesto a actividades de difusión por parte de la universidad. Además, se aprecia que esta cifra viene con una tendencia a la baja, ya que, para el año 2013 el 57% de los matriculados habían tenido actividades de difusión, luego, para el año 2014 esta cifra ya había bajado, llegando al 48%, por lo que en tres años se bajó de un 57% a un 43%, considerando actividades de difusión para alumnos de 4to medio.

Junto con esto se observa que el porcentaje de postulantes que se matricularon y que estuvieron expuestos a difusión sobre el total de convocados fue de solo 23%. Además, la conversión por difusión ha ido disminuyendo en el tiempo, pasando de un 57% en 2014 a un 55% en 2015, como se puede apreciar en la tabla 53 de la sección anexos, de igual forma la cantidad de alumnos de 4to medio que han estado expuesto a difusión disminuyó en un 23,8% entre ambos años. Esto se puede observar en la sección anexos, en la tabla 53 de conversión por difusión.

Otra variable a analizar es la simulación, esto es que la universidad ofrece a los interesados en postular a la universidad un portal en el que puede simular la beca que la casa de estudio le daría si se matricula ahí. Éste monto de la beca es referencial y depende de los puntajes que pueda obtener el alumno, características socioeconómicas y la carrera a la cual postula. El haber realizado esta simulación es requisito para poder optar a una beca de arancel por parte de la institución.

Tabla 12: Conversión por simulación.

	Año 2015			Año 2014			2013		
	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión	Matriculados	Convocados	Conversión
Simula	5.698	9.263	62%	6.886	10.722	64%	5.545	8.851	63%
No simula	1.324	3.953	33%	1.325	4.084	32%	1.675	5.056	33%
Total	**7.022**	**13.216**	**52,3%**	**8.211**	**14.806**	**53,3%**	**7.220**	**13.906**	**48,6%**

Fuente: Elaboración propia, datos Penta Analytics.

En la tabla 12 se observa la conversión que presentan los postulantes que simulan versus los que no simulan en el portal de la universidad. Se aprecia que quienes

realizan esta acción tienen una conversión por sobre el 60%, más específicamente, para el año 2015 la conversión fue del 62% versus un 33% para la conversión de quienes no simularon.

Para el año 2014 y 2013 la situación es similar, quienes simularon a través del portal web tienen una conversión del 64% y 63% en comparación a los 32% y 33% para quienes no simularon respectivamente.

Además, para el año 2015 un 81% de los matriculados simuló beca en la plataforma web, por lo que se puede apreciar que esta es una variable que tiene una alta importancia a la hora de la matrícula, esto se puede observar en la sección anexos en la tabla conversión simulación sobre matriculados.

La cantidad de convocados, en el año 2015, que no cuenta con beca interna al momento de la matrícula es de 6877, de los cuales 2924 a pesar de haber simulado no obtuvieron beneficio para beca de arancel.

En relación al Crédito con Aval del Estado (CAE), la conversión de los postulantes que cuentan con este beneficio ronda el 55% en promedio para los tres años. Por otra parte, la cantidad de matriculados que poseen el beneficio es del 76% para los años 2015 y 2014 y del 70% para el año 2013, esto se puede observar en la sección anexos, en las tablas 55, conversión por CAE, y 56, proporción matriculados con CAE.

En la tabla 56 de anexos se puede ver que la cantidad de los matriculados que posee CAE es sobre el 70% para los tres años, siendo para los años 2014 y 2015 del 76%, por lo que se observa que estos alumnos representan una alta proporción dentro de los matriculados.

Con respecto a la beca externa la conversión es cercana al 60% para los tres años, pero la proporción de alumnos matriculados con beca externa ha ido en aumento a través del tiempo. Entre 2013 y 2014 se experimentó un aumento de un 27% en los matriculados con este beneficio, mientras que entre 2014 y 2015 se aumentó un 12% la cantidad de alumnos que poseían beca externa. Esto hace que la proporción de matriculados con beca externa dentro del total de matriculados este en alza, pasando de un 37% en 2013 a un 54% en 2015. Esto se puede apreciar en anexos en la tabla 57, conversión por beca externa, y 58, proporción matriculados con beca externa.

En relación a la cantidad de alumnos con becas de arancel que otorga directamente la universidad, aumentó en un 21% aproximadamente en relación al año 2014, por lo que hubo un aumento en la cantidad de matriculados con beca, pero por otra parte el beneficio promedio otorgado por la universidad disminuyó en un 19%, por lo que se amplió la cobertura a costa de una reducción en la beca promedio. La cantidad de alumnos que tienen este beneficio representan el 41% del total de matriculados para el año 2015, una cantidad mucho mayor al año 2014 en el cual un 30% de los matriculados contaba con beca interna.

En la tabla 13 se presenta la conversión de las carreras durante el primer período de matrícula para dar cuenta del panorama al cual se enfrenta la universidad, además se presenta la desviación estándar de la conversión de las carreras en los tres años, donde se observa que hay carreras que tienen poca variación y otras que tiene mucha, por lo que predecir el comportamiento de estas carreras es más complicado. Se exhiben algunos programas solamente, la tabla completa se puede ver observar en la sección anexos en las tablas 60-61-62-63.

Tabla 13: Conversión por programa (algunos programas).

Carrera	Sede				Desv. Std.
		2015	2014	2013	
TRABAJO SOCIAL	5	11,6%	13,7%	8,3%	0,04
PSICOPEDAGOGÍA	5	22,2%	27,0%	31,1%	0,03
BACH. HUMANIDADES	2	22,7%	25,0%	38,0%	0,09
TRABAJO SOCIAL	2	26,1%	27,1%	25,3%	0,01
KINESIOLOGÍA	5	26,5%	27,9%	32,4%	0,03
BACH. EN CIENCIAS	2	53,4%	50,8%	51,1%	0,00
MEDICINA VETERINARIA	2	53,6%	53,8%	47,5%	0,04
ECOTURISMO	5	54,1%	48,6%	42,1%	0,05
CONTADOR PUBLICO Y AUDITOR	4	54,2%	35,4%	23,1%	0,09
ING. CIVIL EN COMPUTACION	2	55,2%	54,2%	62,5%	0,06
ENFERMERÍA	4	77,0%	79,4%	81,8%	0,02
GEOLOGÍA	2	77,2%	77,9%	69,3%	0,06
DISEÑO	1	78,4%	66,2%	34,4%	0,22
INGENIERÍA EN BIOTECNOLOGÍA	4	80,0%	84,2%	66,7%	0,12
MEDICINA	4	83,7%	76,0%	85,7%	0,07

Fuente: Elaboración propia, datos Penta Analytics.

En la tabla se observan tres grupos de carreras, ordenadas por conversión según el año 2015, separadas por aquellas que presentan una conversión baja, menor a 30%, conversión media cercana al 50% y las cinco carreras con mayor conversión para el presente año. En esta se observa que las carreras que tienen una conversión baja el año anterior mantiene la tendencia, lo mismo ocurre para aquellas que presentan una conversión media y alta, salvo algunos casos puntuales donde la conversión viene en alza, pero este comportamiento no es la tendencia general de las carreras.

Esto demuestra que la conversión anterior es un buen predictor de cómo le irá a la carrera en el futuro, ya que se mantiene una tendencia.

4.5 FORMULACION DE MODELO

Como se mencionó en la metodología, se desarrollaran cuatro modelos, tres a nivel individual y uno a nivel individual por día. A continuación se presenta el desarrollo para los tres primeros, explicación de sus resultados y la evaluación y validación de los modelos generados.

4.5.1 LOGIT BINOMIAL NIVEL INDIVIDUAL

4.5.1.1 DESARROLLO DEL MODELO

El modelo Logit a nivel individual, logit base, estima la probabilidad de matrícula del postulante en base a sus características, además, busca saber cómo impacta una determinada variable en dicha probabilidad.

Para obtener este modelo se realizaron distintas combinaciones de variables en busca de aquellas que explicaran de forma más precisa la decisión del alumno de matricularse o no.

El modelo utilizado para calcular la probabilidad de matrícula es el siguiente:

$$Pr = \frac{e^{\beta' X}}{1 + e^{\beta' X}}$$

Donde el vector X contiene toda la información del alumno que fue ingresada al modelo. Esta información puede separarse en características propias del alumno, como información demográfica del alumno, información sobre la relación que tuvo el alumno y la universidad e información sobre la carrera a la cual postuló, por lo que el vector de X estaría compuesto de la siguiente forma:

$$X_i = \beta_0 + \beta_1 * Características\ alumno + \beta_2 * Relación\ (Alumno \cap Universidad) + \beta_3 * Información\ carrera$$

Estos tres grandes grupos de información son los que entran al modelo como variables explicativas.

A continuación se detalla las características específicas que contiene cada uno de estos grupos de información.

La información que compone el grupo de características del alumno es la siguiente:

Tabla 14: Características del alumno

Variable	Explicación
Colegio de procedencia	Colegio del cual egresó el alumno, variable binaria por tipo de colegio, particular, subvencionado, municipal.
Quintil	En qué tramo de quintil de ingreso se encuentra el alumno, dejando el quintil 5, el más alto como base.
Año egreso	El año que egresó el alumno, separada en cinco variables binarias, si egresó el mismo año, un año antes, dos años antes, tres años antes y posterior.
Puntaje PSU sobre el promedio de la carrera	Se ocupa el puntaje ponderado del alumno sobre el puntaje promedio de la carrera creando un ratio, estandarizado entre 0 y 1.
Beca externa	Si el alumno tiene beca externa toma el valor 1.
CAE	Si el alumno cuenta con el Crédito con Aval del Estado al momento de la matrícula.
Nivel de estudios del padre	Variable binaria que agrupa niveles de educación del padre del alumno.
Postula a la misma comuna de la sede	Variable binaria que indica si el alumno postuló a una sede de la universidad que queda en su misma comuna.

Fuente: Elaboración propia.

En el vector de características de relación entre el alumno y la universidad se compone de las siguientes variables:

Tabla 15: Variables Relación Alumno-Universidad

Variable	Explicación
Orden de postulación	Variable binaria que indica en qué orden postuló la persona a la universidad.
Simuló misma carrera	Variable binaria que indica si el alumno simuló una beca interna a través del portal web de la universidad a la misma carrera en la que fue convocado.
Log # simulaciones	Variable que indica el logaritmo la cantidad de veces que el postulante simula una beca en el portal de la universidad, estandarizada entre 0 y 1.
Certificado	Variable binaria que indica si el alumno emitió el certificado que valida su postulación para hacer efectiva la beca que le fue ofrecida.
Beca simulada	Esta variable contiene el logaritmo del monto de la beca promedio ofrecida por la universidad al postulante durante la simulación, estandarizada entre 0 y 1.
Difusión	Si el alumno tuvo alguna actividad de difusión por parte de la universidad
Log # exposiciones a difusión	Variable que indica la cantidad de veces que el alumno estuvo expuesto a difusión por parte de la universidad, estandarizada entre 0 y 1
Sede	Variable binaria que indica la sede a la cual postula el alumno.
Diferencia beca simulada y real	Variable que mide la diferencia entre la beca promedio que el alumno simuló y la beca que le corresponde según su puntaje.
Variación beca simulada	Variable que mide la variación en la beca promedio simulada entre un año y el otro para una misma carrera.

Fuente: Elaboración propia.

Finalmente el vector con información de la carrera contiene las siguientes variables:

Tabla 16: Variables Carrera.

Variable	Explicación
Conversión anterior	Variable que indica cual fue la conversión del año recién pasado de la carrera.
Variación carrera en el mercado	Variable que contiene información sobre la variación de la carrera en el mercado universitario. Si se está analizando el año 2015 contiene la variación porcentual de la carrera entre los años 2014 y 2013
Carreras con baja conversión	Variable binaria que indica al 20% de las carreras con la conversión más baja.
Área de conocimiento	Variable que clasifica una carrera con una determinada área del conocimiento.
Valor promedio matrícula IP	Ratio que mide cuanta es la diferencia entre el valor promedio de la matrícula en un Instituto profesional y un programa de la universidad.

Fuente: Elaboración propia.

Para el procesamiento de los datos se utilizó el software estadístico SPSS versión 21, además del software STATA versión 17.

4.5.1.2 RESULTADOS DEL MODELO

En la presente sección se exhiben los resultados obtenidos del modelo descrito anteriormente con las variables de entrada.

En la tabla 17 se muestran las variables introducidas en el modelo con sus respectivos coeficientes, el error estándar, el p-valor y la exponencial del coeficiente (que indica cuánto mejora el modelo al agregar la variable en cuestión). Todas las variables se encuentran escaladas, por lo que los valores de los coeficientes son comparables entre sí.

Para el entrenamiento del modelo se utilizó una base de datos que contenía información de los años 2014 y 2015 para poder tener una calibración más robusta y no dejarse influenciar por uno de los años, ya que el 2014 fue un año muy bueno y el 2015 no lo fue, por lo que ocupar un solo año podía mantener la tendencia del año base.

Tabla 17: Variables introducidas al modelo.

Variable	Coeficiente	Error estándar	P-valor	Exp(B)
Conversión anterior	2,14	0,13	0,00	8,52
Ptje sobre ptje promedio de la carrera	1,43	0,15	0,00	4,18
Postula 1ra preferencia	1,32	0,05	0,00	3,76
Variación entre beca simulada y beca real	1,28	0,33	0,00	3,60
Colegio Particular	0,77	0,06	0,00	2,15
Variación promedio beca simulada	0,62	0,14	0,00	1,85
Egresó un año antes del proceso	0,60	0,07	0,00	1,82
Log # de simulaciones	0,59	0,11	0,00	1,80
Postula 2da preferencia	0,55	0,05	0,00	1,73
Emité certificado	0,50	0,04	0,00	1,65
Variación carrera en el mercado	0,49	0,13	0,00	1,63
Egresó dos años antes del proceso	0,48	0,09	0,00	1,61
Beca externa	0,45	0,04	0,00	1,57
CAE	0,41	0,04	0,00	1,51
Egresó tres años antes del proceso	0,38	0,11	0,00	1,46
Simula misma carrera que postula	0,33	0,05	0,00	1,40
Postula 3ra preferencia	0,31	0,06	0,00	1,36
Egresó mismo año del proceso	0,24	0,07	0,00	1,27
Log beca interna simulada	0,23	0,04	0,00	1,26
Nivel estudios padre universitario	0,19	0,05	0,00	1,21
Colegio Subvencionado	0,19	0,04	0,00	1,21
Postula a una sede de su misma comuna	0,19	0,06	0,00	1,21
Nivel estudios padre CFT-IP	0,17	0,05	0,00	1,19
Nivel estudios padre ed. Media	0,12	0,04	0,00	1,13
Carreras del área salud	0,12	0,04	0,00	1,12
Difusión	0,11	0,04	0,01	1,12
Sede República	-0,08	0,04	0,05	0,93
Sede Casona	-0,22	0,05	0,00	0,80
Sede Bellavista	-0,24	0,07	0,00	0,79
Quintil 4	-0,30	0,06	0,00	0,74
Año 2015	-0,41	0,08	0,00	0,67
Sede Los Leones	-0,41	0,10	0,00	0,66
Quintil 3	-0,52	0,06	0,00	0,60
Quintil 2	-0,74	0,06	0,00	0,48
Quintil 1	-0,85	0,06	0,00	0,43
Carrera con baja conversión y valor matrícula promedio IP	-1,26	0,19	0,00	0,28
Constante	-4,59	0,24	0,00	0,01

Fuente: Elaboración propia.

En la tabla 17 se exhiben los valores de los coeficientes que entrega el modelo para la base de entrenamiento, que se compone del 80% de la base que contiene la información correspondiente a los años 2014 y 2015. Se aprecia que todas las

variables introducidas en el modelo son significativas al 95% de confianza, aún más, la mayoría de las variables lo son al 99%.

En la tabla se aprecia que hay cuatro variables que sobresalen por sobre las otras, que tienen un coeficiente con un valor superior a 1, estas son, en primer lugar, la variable *"Conversión anterior"* que contiene la información pasada de la carrera sobre la ocupación que esta tuvo el año anterior, lo que reafirma la hipótesis de que la historia pasada de la carrera es un buen indicador sobre la conversión futura. Luego le sigue *"Ptje sobre el ptje promedio de la carrera"* que incluye el puntaje ponderado del postulante sobre el puntaje promedio de la carrera, esto indica que si un alumno tiene un buen puntaje tiene mayor probabilidad de matricularse. La tercera variable que se destaca por sobre el resto es *"Postula en 1ra preferencia"* la que contiene la información sobre la preferencia en la postulación de la persona, por lo que hacerlo en primera opción a la universidad aumenta la probabilidad de matrícula de la persona. La última variable que se encuentra en este grupo es *"Variación entre beca simulada y beca real"* que mide la diferencia entre la beca que fue ofrecida a la persona en el simulador, según los datos que ingresó, y la beca que le corresponde según la política de becas para ese año, por lo que si un postulante tiene una beca mayor a la esperada de acuerdo a las simulaciones tiene mayor probabilidad de matrícula.

La interpretación de estas variables es la siguiente. En primer lugar la conversión anterior de la carrera tiene un alto impacto en la decisión del alumno de matricularse o no, y se interpreta que a mayor conversión anterior de la carrera el alumno tiene una mayor probabilidad de matricularse, ya que es una carrera atractiva para los postulantes, por lo que la historia sí importa en la decisión final del postulante. Esta variable entrega aún más información, quiere decir que si a la carrera le fue bien el año pasado este año le debería ir bien, ya que fue una carrera demanda en el pasado, por otra parte, si a la carrera le fue mal anteriormente debería mantener la tendencia, si no existe algún factor externo que modifique la demanda.

Con el puntaje sobre el puntaje promedio de la carrera se puede interpretar de la siguiente manera, mayor diferencia entre el puntaje ponderado del postulante y el puntaje promedio de la carrera este tiene una mayor probabilidad de matricularse en la carrera, esto puede deberse a que a mayor puntaje obtenido en la prueba PSU el alumno tiene una mayor tendencia a matricularse, pues un alto puntaje le permite tener más opciones donde postular, por lo que tiene una mayor probabilidad de elegir la carrera que desea y no postular a la carrera que le "alcanzó", por lo que si postula a la carrera con un buen puntaje quiere decir que realmente está interesado en ella. Además, el tener un buen puntaje le permite optar a mayores beneficios económicos, lo que le ayudaría a solventar los gastos en los que incurre durante el período de estudio.

Una tercera variable que tiene un alto grado de influencia en la decisión del alumno de matricularse o no es la que indica si el alumno postula en primera preferencia a la universidad, esto indica que si el alumno coloca a la institución como primer opción tiene una mayor probabilidad de matricularse, esto debido a que la

postulación indica un orden de preferencia del postulante hacia una casa de estudio, al postular en la primera opción indica que el postulante tiene una mayor intención de elegir la universidad por sobre las otras alternativas. Esta variable tiene un efecto casi tres veces mayor que la variable de segunda preferencia, y más de cuatro veces sobre la tercera. Por lo que demuestra la importancia que tiene postular en la primera alternativa por sobre las otras. Es más, la probabilidad de un alumno que postula en primera preferencia es un 15,6% más alta que la de un postulante que lo hace en segunda opción, y es de un 20,6% frente a alguien que postuló en tercera preferencia. Esta diferencia en la probabilidad de matrícula es aún mayor para alguien que postulo en una preferencia posterior a la 3ra, llegando a 27% más alta.

En cuarto lugar se encuentra la variable diferencia entre beca simulada y beca real, esta variable captura el efecto de que el alumno luego de dar la prueba de selección universitaria y postular a una determinada carrera obtenga un beneficio mayor o menor de los que obtenía en el simulador antes de conocer su puntaje. La variable tiene signo positivo, lo que hace sentido con la hipótesis que se tenía, esto indica que a una mayor variación positiva entre la beca simulada y la real tiene un mayor impacto en la probabilidad de matrícula, y una variación negativa, es decir, que la beca simulada haya sido mayor a la beca que realmente le correspondía tiene un efecto negativo en la probabilidad de matrícula. Que exista una variación entre la beca simulada y la real puede deberse por ejemplo a que el alumno estaba simulando con un puntaje mayor o menor al que obtuvo finalmente en la PSU, por lo que esta diferencia en puntaje lo hizo cambiar de tramo en la política de becas de la universidad y se le asigno la beca de acuerdo a su puntaje real. Otro ejemplo de una variación en la beca podría deberse a que el alumno estaba simulando una carrera a la que finalmente no postuló porque le fue mejor de lo que esperaba y decidió optar por otra carrera o podría ser que le fue peor y el puntaje no le alcanzó para esa carrera y eligió otra opción dentro de la misma universidad.

El colegio de procedencia, es la variable demográfica del postulante que tiene una mayor importancia en la probabilidad de matrícula, en este caso el que tiene una mayor influencia es el Colegio Particular, esto quiere decir que un alumno que egresó de este tipo de establecimiento educacional tiene una probabilidad más alta de matricularse que los postulantes que provienen de otro tipo de colegio, en particular con el postulante que proviene de Colegio Subvencionado donde el coeficiente asociado a esta variable es casi cuatro veces menor. El tipo de colegio que se usó como referencia fue el Municipal. Un alumno que proviene de colegio Particular tiene una probabilidad un 10,8% mayor frente a un alumno de uno Subvencionado y un 14,5% más alta versus a un alumno que proviene de uno Municipal.

A continuación se muestra la tabla 18 con las hipótesis y si nivel de significancia, para validar o rechazar cada una de ellas.

Tabla 18: Validación de hipótesis (1).

Variable	Hipótesis	Significancia
Género	Poder afectar en distinta forma en las ciertas carreras	No Significativa al 90%
Ingreso	A mayor ingreso mayor probabilidad de matrícula	Significativa al 99%
Colegio procedencia	El tipo de colegio puede incidir en la probabilidad matrícula	Significativa al 99%
Estudio del padre	A mayor estudio del padre mayor probabilidad de matrícula	Significativa al 99%
Puntaje PSU	A mayor puntaje aumenta la probabilidad de matrícula	Significativa al 99%
Beca	A mayor beneficio mayor probabilidad de matrícula	Significativa al 99%
Simulación	A mayor simulación aumenta la probabilidad de matrícula	Significativa al 99%
Difusión	A mayor difusión aumenta la probabilidad de matrícula	Significativa al 95%
Certificado	Si emite el certificado en su simulación tiene mayor propensión de matrícula	Significativa al 99%
Cantidad certificados por carrera	A mayor cantidad de certificados emitidos que tenga una carrera mayor probabilidad de conversión	No Significativa al 90%
Año de egreso	A menor tiempo de egreso mayor probabilidad de matrícula	Significativa al 99%
Orden de postulación	Primeras preferencias aumenta la probabilidad de matrícula (1ra, 2da, 3ra)	Significativa al 99%
Otras postulaciones	Disminuye la probabilidad de matrícula	No Significativa al 90%
Carrera con baja conversión y valor matrícula IP's	A mayor diferencia entre el valor de la matrícula de la universidad y de los IP's la probabilidad disminuye	Significativa al 99%
Variación mercado	A mayor crecimiento de la carrera en el mercado mayor probabilidad de matrícula	Significativa al 99%
Variación Universidades Privadas	A mayor crecimiento de la carrera en las universidades privadas mayor probabilidad de matrícula	No Significativa al 90%
Conversión anterior	A mayor conversión anterior mayor probabilidad de matrícula	Significativa al 99%
Valor carrera	A mayor costo de la carrera (ratio arancel de real/referencia) menor probabilidad de matrícula	No Significativa al 90%

Fuente: Elaboración propia.

Tabla 19: Validación de hipótesis (2).

Variable	Hipótesis	Significancia
Comuna sede	Postular a una sede de la misma comuna aumenta la probabilidad de matrícula	Significativa al 99%
Diferencia beca simulada y beca real	A mayor diferencia positiva mayor probabilidad de matrícula	Significativa al 99%
Variación beca simulada	A mayor diferencia positiva mayor probabilidad de matrícula	Significativa al 99%
Variación convocados	A mayor aumento en la cantidad de convocados mayor conversión de la carrera.	No Significativa al 90%
Competencia	Se tienen dos hipótesis contrarias, una es que aumenta la probabilidad de la matrícula y la otra es que disminuye la probabilidad de matrícula.	Significativa al 90%
Área de Conocimiento	La probabilidad de matrícula cambia dependiendo del área de conocimiento a la que pertenezca (área salud)	Significativa al 99%

Fuente: Elaboración propia.

4.5.1.3 CALIDAD DE AJUSTE DEL MODELO

Para medir la calidad de ajuste del modelo se observa los estadísticos definidos anteriormente, comenzando con la tabla de confusión para la base de entrenamiento, calculado con un punto de corte de 0,58.

Tabla 20: Tabla de confusión base entrenamiento.

Observado		Pronosticado		
		matricula		Porcentaje correcto
		No	Si	
matricula	No	7.638	2.643	74,3%
	Si	3.887	8.249	68,0%
Porcentaje global				70,9%

Fuente: Elaboración propia.

Se observa que el modelo clasifica de manera correcta un 70,9% de los casos.

Además, se obtiene la log-verosimilitud que tiene un valor de-12594.8, que nos permite calcular los estimadores como el AIC, BIC y el ratio de verosimilitud.

De la tabla 20 se obtienen los siguientes estadísticos:

Tabla 21: Estadísticos base entrenamiento.

Estadístico	Valor
Log-verosimilitud	-12.594,8
AIC	25.265,6
BIC	25.196,9
Ratio verosimilitud	0,19
Tasa de acierto	70,9%
Tasa de error	29,1%
Recall	68%
Fp rate	25,7%
Precision	75,7%
F-Measure	71,6%

Fuente: Elaboración propia.

Los estadísticos AIC, BIC y ratio de verosimilitud se utilizan para comparar entre dos modelos y determinar cuál es mejor, por lo que serán utilizados para diferenciar con el modelo logit individual con interacciones.

Los otros estadísticos si son para validar el modelo, en particular el estadístico Recall que mide la efectividad para predecir a los postulantes que se matricularon, en este caso el modelo clasifica de manera correcta al 68% de los casos. De igual forma la tasa de aciertos, predecir de manera correcta a quienes se matricularon y no matricularon sobre el total de datos es de un 70,9%, y el estadístico F-Measure, que como se explicó anteriormente es una promedio entre Recall y Precisión, es de 71,6%.

En la tabla 22 se puede observar la probabilidad promedio de los postulantes en cada una de las acciones de marketing (Simulación, Certificado, Difusión, Pre-unab) que puede manejar la universidad según los segmentos de las variables:

1) Tipo de Colegio,
2) Tramo PSU, y
3) Grupo Socioeconómico

En ella se observa que la acción de marketing que tiene una mayor probabilidad de matrícula de los postulantes es la emisión de certificado, sin importar en que variable o segmento de esta se encuentre el postulante, siempre los alumnos que emiten certificado tienen una probabilidad más alta que en otra variable de marketing. Por otra parte, la variable de Pre-unab no fue considerada, ya que en el modelo no era significante y se puede apreciar que la probabilidad de matrícula es muy similar a la de la variable Difusión. Este aspecto es preocupante, ya que para esta actividad (preunab) se destinan aproximadamente $230 millones de pesos al año, lo que equivale a un 22,5% del presupuesto destinado a actividades de Difusión, por lo que aproximadamente un cuarto del presupuesto no está dando los resultados esperados por la universidad, con lo que hay un amplio margen de mejora en este sentido, ya sea redefiniendo las actividades o con una reasignación del presupuesto.

Los segmentos donde se obtiene una mayor probabilidad de matrícula, en cada una de las variables son en alumnos de colegios particulares, en cuanto al tramo PSU (puntaje ponderado) es entre los tramos 700-750 y 750-800, y de acuerdo al grupo socioeconómico es en el quintil de ingreso más alto, el Quintil 5. Por lo que estos segmentos se deben fidelizar e intentar subir la probabilidad de matrícula de los otros segmentos.

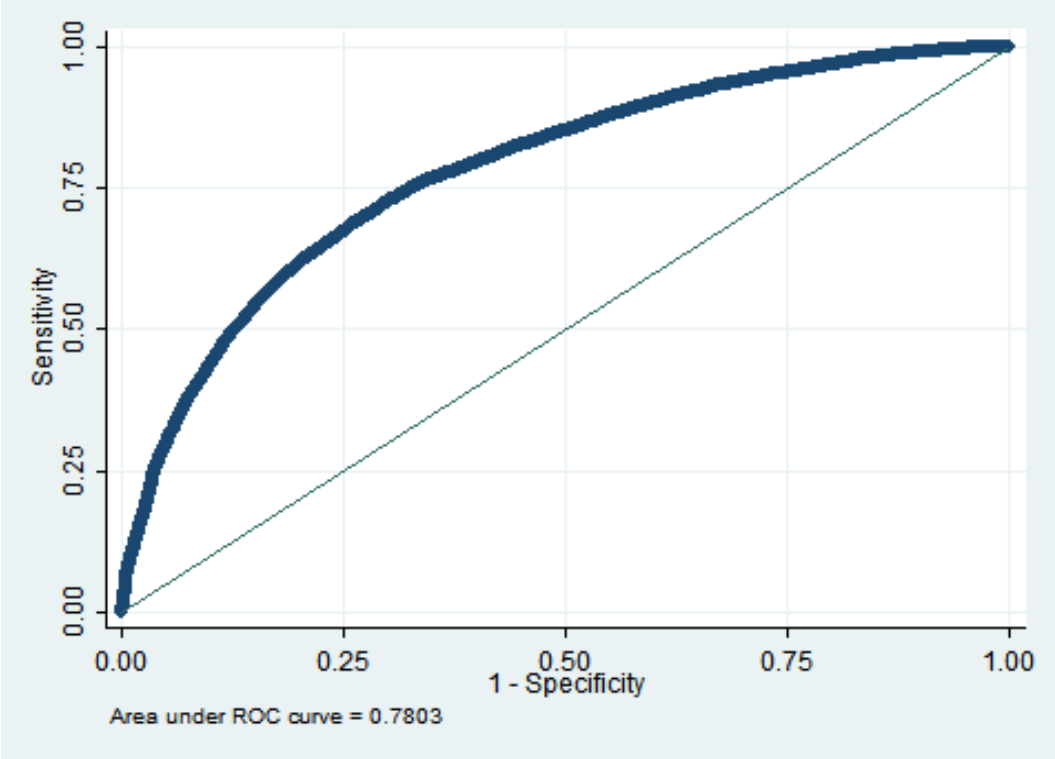
Tabla 22: Acciones Universidad por segmentos.

Variables		**Acciones Universidad**			
	Segmentos	Simulador	Certificado	Difusión	Pre-unab
Tipo de colegio	Particular	77,2%	80,0%	66,6%	65,6%
	Subvencionado	64,7%	68,8%	55,8%	56,0%
	Municipal	58,2%	62,8%	50,6%	50,8%
Tramo PSU	450-500	49,5%	53,7%	39,5%	40,4%
	500-550	60,8%	64,7%	51,6%	51,5%
	550-600	68,8%	72,7%	61,6%	61,5%
	600-650	74,0%	77,7%	67,7%	66,3%
	650-700	75,0%	79,0%	70,7%	70,2%
	700-750	76,4%	79,9%	75,6%	74,6%
	750-800	75,0%	82,2%	72,9%	75,3%
Grupo Socioeconómico	Quintil 1	58,9%	63,3%	50,7%	50,6%
	Quintil 2	62,4%	66,4%	53,1%	53,3%
	Quintil 3	66,9%	70,4%	57,4%	57,3%
	Quintil 4	71,3%	74,9%	62,1%	60,7%
	Quintil 5	75,8%	78,9%	64,6%	64,8%

Fuente: Elaboración propia.

Luego se obtiene la curva ROC del modelo presentado. Esta curva presenta un trade-off entre los beneficios de predecir bien, Recall, y los costos de los falsos positivos, tp-rate.

Gráfico 6: Curva ROC modelo.



Fuente: Elaboración propia.

La línea diagonal muestra un modelo que discrimina entre matriculado y no matriculado como lanzando una moneda, 50% para cada lado, y el modelo se encuentra por sobre esta línea, es decir, el modelo discrimina de mejor forma y es más preciso, abarcando un área bajo la curva del 78%.

Adicionalmente en la tabla 23 se presenta la predicción del modelo en la probabilidad de matrícula de cada alumno. En ella se calcula el porcentaje tanto de los matriculados como de los no matriculados en cada uno de los tramos de probabilidad que predice el modelo.

Tabla 23: Clasificación matriculados y no matriculados base entrenamiento.

Probabilidad	Matriculados	No Matriculados
0%-10%	8%	92%
11%-20%	18%	82%
21%-30%	27%	73%
31%-40%	36%	64%
41%-50%	40%	60%
51%-60%	53%	47%
61%-70%	63%	37%
71%-80%	76%	24%
81%-90%	87%	13%
91%-100%	95%	5%

Fuente: Elaboración propia.

Se observa de la tabla 23 que para una probabilidad de matrícula mayor a 90% el 95% de los alumnos se matricula y solo el 5% de quienes obtienen esta probabilidad de matrícula no lo hacen.

Realizando un análisis exploratorio para los alumnos que pertenecen al 5% que tiene una alta probabilidad de matrícula, pero finalmente no lo hacen se encuentra que estos postulantes presentan características similares, como por ejemplo que el 90% de ellos postula en primera preferencia y el resto postuló en segunda preferencia, que como se comentó anteriormente ambas tiene un efecto positivo en la probabilidad de matrícula. Además, el 74% proviene de colegio particular. Todos ellos simularon un beneficio en el portal web y el 80% de ellos tenía beca e impreso el certificado para validarla. Junto a esto el 52% pertenecía al quintil más alto y el 47% de ellos postula a la sede de Viña del Mar que se ocupó como nivel de referencia en el modelo. Por lo que todas estas características hacen que el modelo les asigne una alta probabilidad de matrícula, pero finalmente ellos no lo hacen.

Realizando un análisis similar al anterior para los postulantes que el modelo le asigna una probabilidad baja, pero igual se matriculan se obtienen características similares, como que solo un 2% postula en primera preferencia, y luego se reparten en las siguientes preferencias, un 29% en la segunda, un 22% en la tercera y el 47% en opciones de preferencia más bajas. Son alumnos que en su mayoría se encuentran por debajo del puntaje promedio de la carrera. El 53% proviene de colegio subvencionado y el resto de colegio municipal, a quienes el modelo le asigna una menor probabilidad de matrícula y solo el 12% de ellos simuló un beneficio en el portal web, y el 5% tuvo beneficio en su simulación, pero ninguno obtuvo su certificado. Por otra parte el 63% pertenece al quintil más bajo de ingreso, que es el que más disminuye la probabilidad de matrícula. Junto a esto el 83% pertenece al grupo del 20% de carreras con más baja conversión, por lo que esto disminuye aún más la probabilidad de matrícula, con lo que se explicaría en parte el por qué el modelo asigna probabilidad baja a estos postulantes, pero se terminan matriculando de todas formas.

4.5.1.4 VALIDACIÓN DEL MODELO

Ahora se calcula la tabla de confusión para la base de validación, que corresponde al 20% de la base 2014-2015, y con ello poder calcular los estadísticos ya mencionados.

A continuación se presenta la tabla de confusión, calculada con el mismo punto de corte de la primera.

Tabla 24: Tabla de confusión base validación.

Observado		Pronosticado		
		matricula		Porcentaje correcto
		No	Si	
matricula	No	1.900	609	75,7%
	Si	956	2.140	69,1%
Porcentaje global				72,1%

Fuente: Elaboración propia.

Se puede observar que los estadísticos para la base de validación son mejores que para la base de entrenamiento, tanto el Recall como el porcentaje global.

Ahora se exhiben los estadísticos para la base de validación.

Tabla 25: Estadísticos base validación.

Estadístico	Valor
Tasa de acierto	72,1%
Tasa de error	27,9%
Recall	69,1%
Fp rate	24,2%
Precision	77,8%
F-Measure	73,1%

Fuente: Elaboración propia.

Los estadísticos para la base de validación son mejores que los obtenidos con la base de entrenamiento, el estadístico Recall es de 69,1%, este es el porcentaje de postulantes que se matriculan y el modelo clasifica de manera correcta y la tasa de aciertos del modelo, es decir, que clasifica de manera correcta el 72,1% de los casos.

A continuación se muestra en la tabla 26 el porcentaje de alumnos matriculados y no matriculados que tienen una determinada probabilidad de matrícula. En la tabla se aprecia que el 86% de los postulantes que tiene más del 90% de probabilidad de matrícula finalmente se matriculan en la universidad y el 14% de ellos no lo hace.

Tabla 26: Clasificación matriculados y no matriculados base validación.

Probabilidad	Matriculados	No Matriculados
0%-10%	12%	88%
11%-20%	15%	85%
21%-30%	28%	72%
31%-40%	35%	65%
41%-50%	39%	61%
51%-60%	56%	44%
61%-70%	66%	34%
71%-80%	78%	22%
81%-90%	87%	13%
91%-100%	93%	7%

Fuente: Elaboración propia.

Realizando el mismo análisis exploratorio anterior, ahora sobre aquel 7% de postulantes que tiene alta probabilidad de matrícula, pero no se matriculan, se observan patrones similares a los encontrados en el grupo de calibración, como por ejemplo que el 92% de ellos postuló en primera preferencia y el resto en la segunda preferencia. Gran parte de ellos, el 65% tiene puntaje ponderado sobre el promedio de la carrera. El 58% proviene de colegio particular y el 100% simuló un beneficio y todos obtuvieron beca, y el 92% de ellos emitió el certificado para validar la beca y el 50% aproximadamente pertenece al quintil 5.

En relación al 12% de postulantes que el modelo le asigna baja probabilidad de matrícula, pero que de igual forma se matriculan se puede decir que estos tienen en común las siguientes características, como lo son que ninguno postula en primera preferencia, un 25% lo hace en segunda y otro 25% en tercera preferencia, el 50% restante lo realiza desde la cuarta opción en adelante. Un 37,5% de los postulantes realiza simulaciones, pero un 12,5% de estos simula la misma carrera a la que fue convocado y de estos solo el 6% obtiene una beca. Un 56,2% pertenece al primer quintil de ingreso y un 31,2% al segundo quintil. Y un 69% pertenece al grupo de carreras con conversión más baja. Por lo tanto, todas estas características que tienen en común los postulantes provoca que el modelo le asigne una baja probabilidad de a pesar de que se matriculan en la universidad.

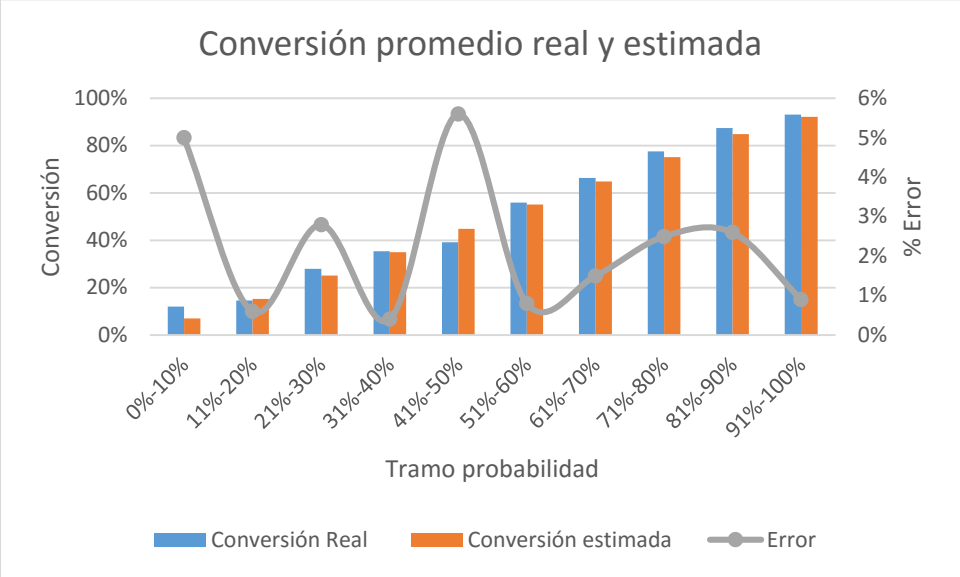
También se presenta, en el gráfico 7, la conversión real y estimada por el modelo por tramo de probabilidad.

En él se puede observar que la conversión real y la probabilidad promedio de matrícula estimada por el modelo es muy similar, presentando una diferencia máxima de un 5,6% para el tramo 41%-50% donde se sobreestima la probabilidad de matrícula para ese tramo y en el tramo de 0%-10% que la diferencia es de 5%, pero en este segmento se subestima la probabilidad. Para los demás tramos la diferencia es menor y en algunos casos menor al 1%.

El error promedio en la conversión es de 2,27%, por lo que el modelo estima de manera correcta la conversión y es muy similar a la real, como se puede observar en

el gráfico 7. En la sección anexos se puede encontrar la tabla con la cantidad de postulantes por grupo, además de la conversión y probabilidad estimada.

Gráfico 7: Conversión real y estimada por el modelo Logit.



Fuente: Elaboración propia.

4.5.2 LOGIT BINOMIAL NIVEL INDIVIDUAL CON INTERACCIONES

4.5.2.1 DESARROLLO DEL MODELO

Al igual que en el modelo anterior este tiene por finalidad estimar la probabilidad de matrícula de un convocado seleccionado, por lo que tiene la misma base de variables. La variación de este modelo con respecto al anterior es que contiene interacciones entre variables, para dar un ejemplo, se agrega interacción entre el tipo de colegio de procedencia del postulante y la beca ofrecida por la universidad.

El modelo utilizado es el mismo que el utilizado anteriormente y viene dado por la siguiente expresión.

$$Pr = \frac{e^{\beta' X}}{1 + e^{\beta' X}}$$

Las variables utilizadas en este modelo pueden ser clasificadas de la siguiente forma:

$$X_i = \beta_0 + \beta_1 * Características\ alumno + \beta_2 * Relación\ (Alumno \cap Universidad) + \beta_3 * Información\ carrera + \beta_4 * Interacciones$$

A continuación se explica cada una de las variables que se introdujeron al modelo. Se puede apreciar que en las características del alumno presenta una menor cantidad de variables, esto se debe a que al ingresar algunas de estas variables como interacciones no se tienen que agregar por si solas, pues genera distorsión en el modelo y provoca que la nueva variable no sea significativa, por lo que se decide introducir solamente como una interacción para estudiar ese efecto en particular, ya que el otro efecto por si solo fue estudiado en el modelo anterior.

Tabla 27: Características del alumno

Variable	Explicación
Año egreso	El año que egresó el alumno, separada en cinco variables binarias, si egresó el mismo año, un año antes, dos años antes, tres años antes y posterior.
Puntaje PSU sobre el promedio de la carrera	Se ocupa el puntaje ponderado del alumno sobre el puntaje promedio de la carrera creando un ratio, estandarizado entre 0 y 1.
Beca externa	Si el alumno tiene beca externa toma el valor 1.
CAE	Si el alumno cuenta con el Crédito con Aval del Estado al momento de la matrícula.
Postula a la misma comuna de la sede	Variable binaria que indica si el alumno postuló a una sede de la universidad que queda en su misma comuna.

Fuente: Elaboración propia.

En el vector de características de relación entre el alumno y la universidad se compone de las siguientes variables:

Tabla 28: Variables Relación Alumno-Universidad

Variable	Explicación
Orden de postulación	Variable binaria que indica en qué orden postuló la persona a la universidad.
Simuló misma carrera	Variable binaria que indica si el alumno simuló una beca interna a través del portal web de la universidad a la misma carrera en la que fue convocado.
Log # simulaciones	Variable que indica el logaritmo la cantidad de veces que el postulante simula una beca en el portal de la universidad, estandarizada entre 0 y 1.
Certificado	Variable binaria que indica si el alumno emitió el certificado que valida su postulación para hacer efectiva la beca que le fue ofrecida.
Log # exposiciones a difusión	Variable que indica la cantidad de veces que el alumno estuvo expuesto a difusión por parte de la universidad, estandarizada entre 0 y 1
Diferencia beca simulada y real	Variable que mide la diferencia entre la beca promedio que el alumno simuló y la beca que le corresponde según su puntaje.
Variación beca simulada	Variable que mide la variación en la beca promedio simulada entre un año y el otro para una misma carrera.

Fuente: Elaboración propia.

El vector con información de la carrera contiene las mismas variables que el modelo anterior, de igual forma se muestran en la tabla.

Tabla 29: Variables Carrera.

Variable	Explicación
Conversión anterior	Variable que indica cuál fue la conversión del año recién pasado de la carrera.
Variación carrera en el mercado	Variable que contiene información sobre la variación de la carrera en el mercado universitario. Si se está analizando el año 2015 contiene la variación porcentual de la carrera entre los años 2014 y 2013
Área de conocimiento	Variable que clasifica una carrera con una determinada área del conocimiento

Fuente: Elaboración propia.

En el vector de interacciones se agregan las siguientes variables:

Tabla 30: Variables Interacción.

Variable	Explicación
Beca Interna-Tipo Colegio	Variable que contiene el monto de la beca simulada y el tipo de colegio de precedencia del alumno.
Sede-Quintil	Las sedes se pueden ver afectada de distintas formas según el nivel de ingreso del alumno.
Difusión-Tipo colegio	Variable que contiene información sobre a qué colegio le es más efectiva la difusión por parte de la universidad.
Variación de la carrera en el mercado-Variación en los convocados	Variable que mide el efecto de la variación de la carrera en el mercado y su influencia en la variación de los convocados a una determinada carrera.
Nivel estudio padre-Beca Interna	Los postulantes pueden ser más sensibles a la beca dependiendo del nivel de estudios del padre, ya que un cierto nivel de estudio supone un cierto nivel de ingreso, por eso se pretende medir el efecto en cada caso.
Nivel competencia-Sede	El nivel de competencia puede influir de diferente forma dependiendo de las sedes, ya que cada sede ofrece distintas carrera y cada una de ellas tiene un nivel de competencia.
Carreras baja conversión- Valor promedio matricula IP	Al igual que en el modelo anterior, las carreras con baja conversión puede deberse a que no son muy atractivas y si el valor de la matrícula es mucho mayor que el de un IP, los postulantes a estas carreras prefieren el IP ya que es menos tiempo de estudio y la inversión es menor.

Fuente: Elaboración propia.

4.5.2.2 RESULTADOS DEL MODELO

Los coeficientes que entrega el modelo se presentan en la tabla que se despliega a continuación.

Al igual que en el modelo anterior la tabla contiene los coeficientes de las variables, la desviación estándar, el P-valor y el exp(B).

Dado que se cuenta con muchas variables, debido a la cantidad de interacciones entre la sede y los quintiles en mayor medida se exponen los resultados en dos tablas, donde la tabla 3 es la continuación a la tabla 31.

Tabla 31: Coeficientes modelo con interacciones (1ra parte).

Variable	Coeficiente	Error estándar	P-valor	Exp(B)
Conversión anterior	2,09	0,13	0,00	8,11
Variación entre beca simulada y beca real	1,75	0,14	0,00	5,74
Ptje sobre ptje promedio de la carrera	1,39	0,33	0,00	4,01
Postula 1ra preferencia	1,30	0,05	0,00	3,68
Egresó un año antes del proceso	0,71	0,07	0,00	2,04
Colegio Particular y Difusión	0,64	0,07	0,00	1,90
Log # de simulaciones	0,57	0,11	0,00	1,76
Colegio Particular y Beca interna	0,56	0,09	0,00	1,75
Postula 2da preferencia	0,55	0,05	0,00	1,73
Egresó dos años antes del proceso	0,54	0,09	0,00	1,71
Variación promedio beca simulada	0,53	0,14	0,00	1,71
Variación de convocados y Variación de mercado	0,53	0,21	0,01	1,70
# Certificados emitidos por carrera	0,52	0,10	0,00	1,68
Emité certificado	0,51	0,04	0,00	1,67
Egresó tres años antes del proceso	0,38	0,11	0,00	1,47
Egresó mismo año del proceso	0,37	0,07	0,00	1,44
CAE	0,37	0,04	0,00	1,44
Postula 3ra preferencia	0,31	0,06	0,00	1,36
Simula misma carrera que postula	0,31	0,05	0,00	1,36
Colegio Subvencionado y Beca externa	0,27	0,04	0,00	1,32
Postula a una sede de su misma comuna	0,21	0,06	0,00	1,24

Fuente: Elaboración propia.

Tabla 32: Coeficientes modelo con interacciones (2da parte).

Variable	Coeficiente	Error estándar	P-valor	Exp(B)
# Exposiciones de difusión	0,18	0,08	0,03	1,20
Nivel estudio padre Ed. Universitaria y Beca interna	0,15	0,07	0,03	1,16
Nivel estudio padre Ed. Media y Beca interna	0,15	0,06	0,01	1,16
Género masculino	0,13	0,03	0,00	1,14
Colegio Subvencionado y Beca interna	0,11	0,06	0,04	1,12
Carreras del área salud	0,07	0,04	0,08	1,07
Quintil 3 y Sede Repúbllica	-0,14	0,07	0,04	0,87
Quintil 3 y Sede Casona	-0,25	0,08	0,00	0,78
Año 2015	-0,33	0,08	0,00	0,72
Quintil 2 y Sede República	-0,33	0,07	0,00	0,72
Nivel de competencia y Sede Bellavista	-0,39	0,09	0,00	0,68
Quintil 1 y Sede Bellavista	-0,39	0,12	0,00	0,67
Quintil 1 y Sede República	-0,46	0,05	0,00	0,63
Quintil 3 y Sede Los Leones	-0,53	0,23	0,02	0,59
Quintil 2 y Sede Casona	-0,59	0,08	0,00	0,55
Quintil 1 y Sede Casona	-0,75	0,07	0,00	0,47
Quintil 1 y Sede Los Leones	-0,77	0,15	0,00	0,46
Quintil2 y Sede Los Leones	-0,86	0,21	0,00	0,43
Nivel de competencia y Sede Concepción	-0,96	0,23	0,00	0,38
Carrera con baja conversión y valor matrícula promedio IP	-1,21	0,19	0,00	0,30
Constante	-4,91	0,23	0,00	0,01

Fuente: Elaboración propia.

De la tabla 31 y 32, de coeficientes, se aprecia que, al igual que en el modelo anterior, las variables que tienen una mayor importancia son las mismas, en primer lugar la *Conversión anterior*, y solo se presenta un cambio de orden de la *Variación de la beca simula y real*, que sube a la segunda posición desplazando a las variables de *Puntaje ponderado sobre el promedio* y la *Postulación en primera preferencia*. De igual forma, estas cuatro variables mantienen un coeficiente mayor a 1 la conversión anterior de la carrera. Además, todas las variables son significativas, como mínimo al 90% donde hay una variable en este nivel de significancia. El resto todas significativas al 95%, incluso algunas de ellas al 99%.

El modelo cuenta con variables que se mantienen sin variación del modelo anterior, solo presenta una variación en el valor del coeficiente, pero la interpretación es la misma, por lo que este grupo de variables no serán explicadas, ya que este análisis fue realizado en el primer modelo. Las variables que conforman este grupo son:

- Conversión anterior
- Variación entre la beca simulada y la real
- Puntaje ponderado sobre el promedio de la carrera

- Preferencia de postulación del alumno
- Año de egreso del postulante
- Cantidad de simulaciones
- Variación promedio beca simulada
- Emisión de certificado por parte del postulante
- CAE
- Simula la misma carrera que postula, y
- Carreras del área de la salud, Año 2015, Carreras con baja conversión y precio promedio de la matrícula de los Institutos Profesionales.

A continuación se explicarán las variables nuevas e interacciones que fueron introducidas al modelo.

La primera variable con interacciones que aparece es aquella que agrupa la información sobre el colegio de procedencia y la difusión que realiza la universidad. En este caso, el modelo indica que la difusión es más efectiva para los postulantes que provienen de colegios particulares por sobre los que provienen de colegios subvencionados o municipales. Es más, esta variable no es significativa para los postulantes que provienen de colegios subvencionados, por lo que no aparece en el modelo. Es por esto que si un alumno proviene de colegio particular y estuvo expuesto a difusión es más propenso a matricularse que uno que viene de otro tipo de colegio que tuvo difusión.

La siguiente variable que aparece también tiene relación con el tipo de colegio de procedencia y la beca interna que otorga la universidad. Esta es una relación positiva, por lo que a mayor beca que le ofrece la universidad a un alumno de este tipo de colegio mayor probabilidad de matrícula. Incluso esta relación entre la beca interna y el alumno procedente de colegio particular es mayor que para los alumnos que egresaron de un colegio subvencionado y municipal. La diferencia entre las magnitudes de los coeficientes para esta variable entre los alumnos de colegio particular y subvencionado es de cinco veces mayor. Por lo que, al tener una mayor sensibilidad a la beca conviene darle beca a un alumno de colegio particular por sobre postulante de otros colegios. Un alumno que proviene de colegio particular y cuenta con beca interna de arancel tiene una probabilidad un 10% más alta que si el mismo alumno proviene de un colegio subvencionado con beca interna de arancel, y tienen una probabilidad un 15% mayor en caso de que provinieran de colegio particular y no tuvieran esta beca.

Le sigue la variable que mide el efecto en la interacción entre la variación de la carrera en el mercado y la variación de los convocados en una carrera. Este efecto es positivo, por lo que si la ésta tiene una variación positiva en el mercado, es decir, que está siendo cotizada y demanda por los alumnos, la carrera va a tener una mayor convocatoria que el año anterior, y como la carrera está siendo más demandada y

hay una mayor convocatoria la probabilidad de que el postulante se matricule es mayor.

Otra variable que fue incluida en este modelo es la cantidad de certificados emitidos por carrera, variable que también es positiva, por lo que a un mayor número de certificados emitidos quiere decir que la carrera está siendo más simulada (demandada) y que a los postulantes les interesa el programa por eso quieren el certificado para luego poder validar el beneficio ofrecido.

Otra variable que fue introducida al modelo y que no se encontraba en el anterior fue la cantidad de exposiciones a difusión que tuvo el postulante, como el signo es positivo quiere decir que a mayor cantidad de exposiciones a difusión la probabilidad de matrícula aumenta, por lo que la difusión si tiene efecto en los postulantes, aunque el valor del coeficiente no es muy alto. Además, se incluyó interacción entre tipo de colegio y difusión, concluyendo que esta es significativa para postulantes que provienen de colegio particular por sobre los colegios subvencionados y municipales. Un alumno de colegio particular que estuvo expuesto a estas actividades tiene una probabilidad un 11,8% más alta frente a un alumno de colegio particular que no tuvo difusión.

Las siguientes variables con interacciones son la que agrupa la información sobre el nivel de estudios del padre y la beca ofrecida por la universidad, esta variable indica que para postulantes que tienen padre con estudio universitario y enseñanza media la beca interna es importante y significativa, tiene un efecto levemente superior para quienes tienen padres con estudios universitarios, pero la diferencia está en el tercer decimal, por lo que no hay mucha diferencia entre uno y otro, pero en ambos postulantes tiene un efecto positivo la beca interna.

Otra variable que se agregó fue el género del postulantes, en este caso si el postulante es de género masculino tiene una probabilidad mayor que un postulante de género femenino, aunque el valor del coeficiente no es muy alto, por lo que la diferencia entre la probabilidad de matrícula es un poco mayo entre un postulante de género masculino y femenino.

Luego aparecen las variables que relacionan el nivel de ingreso de los postulantes y las sedes a las que postulan. Con estas interacciones solo se tuvo variables significativas para los primeros tres quintiles, para el cuarto las interacciones no eran significativas, por lo que no ingresaron al modelo. Situación similar ocurrió con las sedes, solo quedaron tres que tuvieron interacción significante con el quintil dos y tres, y cuatro sedes con el primer quintil. Todas estas interacciones son negativas, por lo que disminuye la probabilidad de matrícula del postulante, pero lo que es interesante es que el quintil afecta de manera distinta a cada sede. Se tiene que la sede que tiene un menor impacto negativo para los postulantes que pertenecen al quintil tres es República y la mayor es la sede Los Leones, con una diferencia casi cuatro veces mayor. De hecho la sede Los Leones es la que se ve mayormente afectada en la interacción con los quintiles, siempre es la que tiene el coeficiente más negativo dentro de un mismo quintil. De hecho, un alumno que postula a la sede los

leones y pertenece al quintil 1 tiene una probabilidad un 15,6% más baja frente a que si postulara a la sede de Viña o de Concepción, y una probabilidad un 6,6% más baja en caso de que postulara a la sede de República. Esta brecha se va reduciendo a medida que aumenta en el tramo de ingreso, ya que si un postulante del quintil 2 postula a la sede Los Leones tiene una probabilidad un 13% más baja frente a la misma persona que postula a las sedes de Viña o Concepción, y para el tercer quintil de ingreso se reduce aún más, llegando a una diferencia de 10,6% entre una persona que postula a Los Leones frente a que si postula a Viña o Concepción.

La última variable que fue incluida en el modelo es la que mide el nivel de competencia de las carreras y la sede, en esta variable solo quedaron dos de las cinco sedes (una sede se mantiene como nivel de referencia). En ambos casos el coeficiente es negativo, por lo que a una mayor cantidad de universidades que ofrezcan el programa afecta de manera negativa la probabilidad de matrícula de los postulantes. Las sedes que quedaron en el modelo fueron Bellavista y Concepción, siendo esta última la que tiene un mayor efecto negativo en la probabilidad, el coeficiente aproximadamente 2,5 veces mayor que para la sede Bellavista, por lo que esta sede siente en mayor medida la competencia que las otras sedes de la universidad.

La calidad de ajuste del modelo logit con interacciones se podrá ver en anexos, ya que es muy similar a la realizada para el modelo anterior.

4.5.2.3 VALIDACIÓN DEL MODELO

A continuación se presenta la tabla de confusión para la base de validación, que se compone del 20% de la base que contiene la información para el 2015-2014.

Tabla 33: Tabla de confusión base validación.

Observado		**Pronosticado**		
		matricula		Porcentaje correcto
		No	Si	
matricula	No	1.902	607	75,8%
	Si	1047	2.049	66,2%
Porcentaje global				70,5%

Fuente: Elaboración propia.

Ahora se exhiben los estadísticos para la base de validación.

Tabla 34: Estadísticos base validación.

Estadístico	Valor
Tasa de acierto	71,2%
Tasa de error	28,8%
Recall	73,2%
Fp rate	31,5%
Precision	74,5%
F-Measure	73,9%

Fuente: Elaboración propia.

De los estadísticos se observa que la base de validación tiene un peor ajuste que la base de entrenamiento, en el estadístico Recall, es decir, que tiene un menor poder de detección de los alumnos que si se matricularon, por otra parte el porcentaje correcto de detección de aquellos que no se matricularon es mayor para esta base que para la de entrenamiento, al igual que el porcentaje global de aciertos, que es mayor para esta base.

Mencionar que tanto la base de entrenamiento como la de validación es la misma para ambos modelos.

A continuación se presenta el porcentaje de alumnos según la probabilidad de matrícula que el modelo le asigna a cada uno.

Tabla 35: Clasificación matriculados y no matriculados base validación.

Probabilidad	Matriculados	No Matriculados
0%-10%	12%	88%
11%-20%	18%	82%
21%-30%	28%	72%
31%-40%	37%	63%
41%-50%	46%	54%
51%-60%	59%	41%
61%-70%	67%	33%
71%-80%	78%	22%
81%-90%	90%	10%
91%-100%	91%	9%

Fuente: Elaboración propia.

Se aprecia que para la base de validación el modelo es menos preciso, ya que le asigna a un mayor grupo una probabilidad baja de matrícula que si se matriculan y de igual forma a un mayor grupo una probabilidad alta de matrícula a quienes no lo terminan haciendo finalmente, en comparación con la base de entrenamiento.

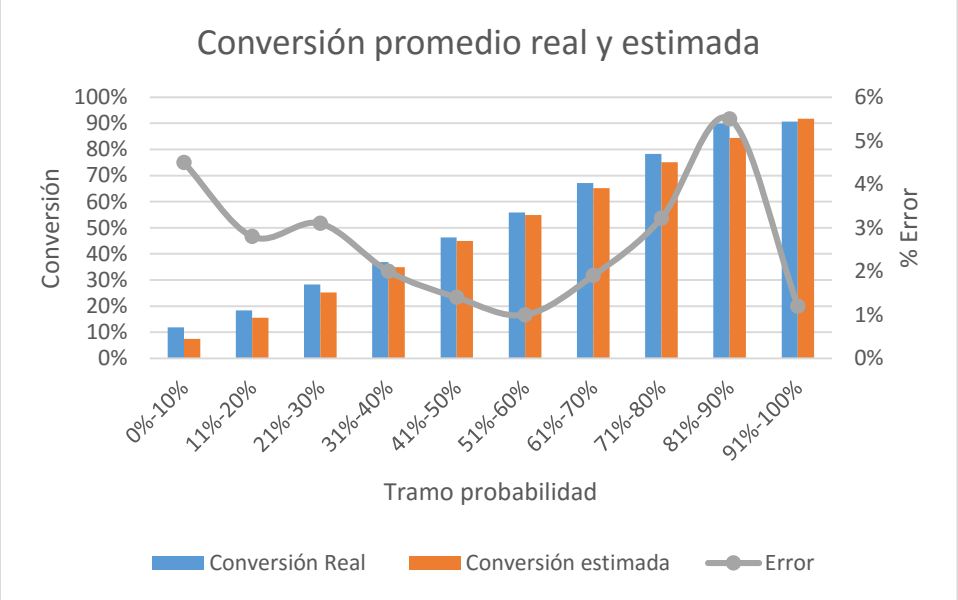
Realizando el mismo análisis anterior para identificar las características comunes de los postulantes en estos grupos se observa que para aquellos postulantes a los que el modelo le asigna una alta probabilidad, pero que no se matriculan se caracterizan por postular en primera preferencia, el 91% de ellos lo realizó y el 9% lo hizo en

segunda opción. En general son alumnos que tienen puntaje ponderado sobre el promedio y el 100% de ellos realizó simulación de beneficios en el portal web, y todos ellos lo hicieron para la misma carrera en la que fue convocado, y todos ellos obtuvieron beca a través de su simulación y emitieron el certificado correspondiente para validarla. Un 64% pertenece a los dos quintiles de ingreso más alto y postulan a carreras que tiene una alta conversión, por estos motivos el modelo sobreestima su probabilidad de matrícula siendo que finalmente no se matriculan.

En relación al grupo de baja probabilidad que sí se matricula, las variables que caracterizan a este grupo son que el 64% de ellos postula a la universidad después de la tercera preferencia, un 28% de ellos realiza simulaciones, pero sólo un 7% lo hace para la misma carrera a la que es convocado y solo ese 7% obtiene una beca interna producto de la simulación y el 86% del grupo pertenece a los dos primeros quintiles de más bajo ingreso, junto a esto el 78% de estos alumnos postulan al grupo del 20% de carreras con más baja conversión, por lo que todas estas características hacen que se les asigne una baja probabilidad de matrícula siendo que si lo hacen.

En el gráfico 8 se observa la conversión real y estimada por el modelo por tramo de probabilidad. Se aprecia que la conversión promedio es muy similar a la conversión real, y se presenta un error mayor en los tramos 1%-10%, al igual que en el modelo logit base, y 81%-90% donde se alcanza el máximo error de 5,5%. El error promedio es de 2,66% mayor que el modelo anterior que es de 2,27%, por lo que tiene un peor ajuste en comparación al modelo sin interacciones. En la sección anexos se puede encontrar la tabla 60 con los porcentajes y cantidad de postulantes por tramo.

***Gráfico* 8:** Conversión real y estimada por el modelo Logit con interacciones.



Fuente: Elaboración propia.

4.5.3 ÁRBOL DE DECISIÓN

4.5.3.1 DESARROLLO DEL MODELO

Para el desarrollo de este modelo se usaron todas las variables que se tenían disponibles para que el mismo árbol discriminara cuáles eran importantes y quedaban en el modelo y cuáles no, por lo que sólo se explicarán las variables nuevas que le fueron introducidas en él y que no hayan sido explicadas en los modelos anteriores.

Para la construcción del árbol de decisión no se utilizarán variables con interacciones, ya que este modelo tenía un ajuste un poco menor que el modelo sin estas relaciones entre variables, por lo que se quiere comparar el mejor modelo versus un árbol de decisión. El método utilizado para el crecimiento del árbol es método CHAID.

Las variables que se ocuparon para realizar el modelo, y que ya fueron explicadas son las siguientes:

- Simula misma carrera
- 1ra preferencia
- Carreras pertenecientes al 20% de más baja conversión
- Quintil 1
- Variación de la carrera en el mercado
- Año 2015
- Sede 1
- Si emite certificado
- Puntaje ponderado sobre el promedio

Las nuevas variables que fueron introducidas al modelo y que finalmente quedaron en él son:

- Variación de la universidad
- Ratio arancel sobre las universidades

Tabla 36: Variables nuevas introducidas al árbol de decisión

Variable	Explicación
Variación de la carrera a nivel universidad	Variable que mide la variación de la carrera a nivel de la universidad, para saber si esta viene a la baja o al alza dentro de la universidad.
Arancel real de la carrera sobre el arancel real promedio de las universidades del sistema	Variable continua que mide cuanto más es el valor del arancel real versus el arancel real promedio de las universidades que ofrecen el mismo programa.

Fuente: Elaboración propia.

4.5.3.2 RESULTADOS DEL MODELO

Para observar los resultados del modelo sólo se mostrará el camino desde el nodo raíz al mejor nodo, a través de una ilustración, debido al tamaño del árbol.

Para generar el modelo se ocupa el 80% de la base que contiene información de los años 2015 y 2014.

El resultado del árbol se muestra de la siguiente forma:

Ilustración 2: Resultado árbol de decisión.

Variable y valor		
Número del nodo		
Categoría	Porcentaje	Cantidad
0	45,7%	10272
1	54,3%	12218
Total	100%	22490

Fuente: Elaboración propia.

Donde en la primera casilla se muestra la variable y el valor que toma en el nodo. Luego se observa el número del nodo. Categoría se divide en dos {0,1}, donde 1 indica que el alumno se matricula y 0 que no lo hace, de igual forma, muestra qué porcentaje y la cantidad de los postulantes que lo hace o no, dentro de los postulantes que tiene información en esa variable. Como se ve a continuación (el diagrama del árbol se puede ver completo en el anexo 11, para ver cuáles son los nodos por los que se llega al nodo final).

Ilustración 3: Resumen árbol de decisión.

Matrícula		
Nodo 0		
Categoría	Porcentaje	Cantidad
0	45,7%	10272
1	54,3%	12218
Total	100%	22490

Simula misma carrera=1		
Nodo 2		
Categoría	Porcentaje	Cantidad
0	34,8%	5236
1	65,2%	9806
Total	67%	15042

1ra preferencia=1		
Nodo 6		
Categoría	Porcentaje	Cantidad
0	25,8%	2558
1	74,2%	7358
Total	44%	9916

Emite Certificado=1		
Nodo 12		
Categoría	Porcentaje	Cantidad
0	22,5%	1799
1	77,5%	6201
Total	36%	8000

Ptje sobre el promedio>0,40		
Nodo 26		
Categoría	Porcentaje	Cantidad
0	15,9%	314
1	84,1%	1663
Total	9%	1977

Fuente: Elaboración propia.

Entonces las variables que importan a la hora de la matrícula es: Si simula en la misma carrera, si postula en primera preferencia, si emite certificado, y si tiene un puntaje ponderado sobre el promedio mayor a 0,4. Esta última variable se encuentra estandarizada, por lo que un puntaje ponderado sobre el promedio de 0,4 significa que la diferencia entre el puntaje del postulante y del promedio de la carrera no sea inferior a 90 puntos.

Se observa que los postulantes que reúnen estas características tienen la probabilidad de matrícula del 86% y el 14% restante no se matricula, en total son 1452 postulantes que tienen estas características y que se matriculan finalmente

4.5.3.3 CALIDAD DE AJUSTE DEL MODELO

La calidad de ajuste del modelo será medida por la tabla de confusión para poder comparar con los modelos desarrollados anteriormente.

Tabla 37: Tabla de confusión base entrenamiento.

Observado		**Pronosticado**		
		matricula		Porcentaje correcto
		No	Si	
matricula	No	6.282	3.990	61,2%
	Si	3.130	9.088	74,4%
Porcentaje global				68,3%

Fuente: Elaboración propia

Y a continuación se presenta la tabla con los estadísticos calculados para el modelo.

Tabla 38: Estadísticos base entrenamiento.

Estadístico	Valor
Tasa de acierto	68,3%
Tasa de error	31,7%
Recall	74,4%
Fp rate	38,8%
Precision	69,4%
F-Measure	71,8%

Fuente: Elaboración propia

Los estadísticos para el modelo son levemente peores que para los modelos logit desarrollados anteriormente. De igual forma, el estadístico Recall es superior al 70%

4.5.3.4 VALIDACIÓN DEL MODELO

Para la validación del modelo se ocupó el 20% de la base que contiene información para los años 2015 y 2014. En esta ocasión el mismo modelo separa los datos para calibración y validación.

Para la base de validación también se exhibe la tabla de confusión.

Tabla 39: Tabla de confusión base entrenamiento.

Observado		**Pronosticado**		
		matricula		Porcentaje correcto
		No	Si	
matricula	No	1.554	964	61,7%
	Si	794	2.220	73,7
Porcentaje global				68,2%

Fuente: Elaboración propia

Y a continuación se presenta la tabla 40 con los estadísticos calculados para el modelo.

Tabla 40: Estadísticos base entrenamiento.

Estadístico	Valor
Tasa de acierto	68,2%
Tasa de error	31,8%
Recall	73,7%
Fp rate	38,2%
Precision	69,7%
F-Measure	71,6%

Fuente: Elaboración propia

Al igual que para los datos de validación en los modelos anteriores, los estadísticos tienen una menor precisión que para la base de calibración. Sin embargo, se siguen manteniendo cerca del 70% el estadístico Recall, Precision y el F-Measure.

4.6 COMPARACIÓN DE LOS MODELOS

En esta sección se harán dos comparaciones, en primer lugar se compararán los tres modelos generados, el logit base, el modelo logit con interacciones y el árbol de decisión. Luego se realizará una comparación entre los modelos que tengan un mejor ajuste y predicción de la primera comparación y el desarrollado por la empresa sponsor Penta Analytics, que cuenta con un modelo Logit para la estimación de matrícula.

4.6.1.1 COMPARACIÓN MODELOS LOGIT Y ÁRBOL DE DECISIÓN

Para comparar estos modelos se utilizarán los estadísticos obtenidos de la tabla de confusión, con punto de corte 0.5, tanto para la base de calibración y de validación.

En primer lugar se compararán los estadísticos para la base de entrenamiento de los modelos.

Tabla 41: Comparación base de entrenamiento.

Estadístico	Logit base	Logit con interacciones	Árbol de decisión
Tasa de acierto	**72,6%**	71,2%	68,3%
Tasa de error	**27,4%**	28,8%	31,7%
Recall mat.	**77,6%**	73,2%	74,4%
Recall no mat.	66,5%	64,6%	**66,7%**
Fp rate	33,4%	**31,5%**	38,8%
Precision mat.	74,1%	**74,5%**	69,4%
Precision no mat	**70,6%**	69,6%	66,7%
F-Measure	**75,8%**	73,9%	71,8%
Estimación matrícula	12.909	12.877	13.078

Fuente: Elaboración propia

Como se puede observar en la tabla de comparación los mejores estadísticos se encuentran en el modelo logit base, luego le sigue el modelo logit con interacciones y en tercer lugar y último se encuentra el árbol de decisión, que tiene solamente un estadístico mejor que los modelos logit, que son el Recall de los no matriculados, que es levemente mejores en comparación a los modelos logit.

Ahora comparando ambos modelos logit se observa que el modelo base tiene todos sus estadísticos mejores que el modelo logit con interacciones, si bien, no se observa grandes diferencias entre un estimador y otro, estos si son mejores en el ajuste a los datos, en este caso se utiliza los mismos datos para la calibración.

A continuación se exhibe la tabla 42 de comparación de los estadísticos para la base de validación.

Tabla 42: Comparación base de entrenamiento.

Estadístico	Logit base	Logit con interacciones	Árbol de decisión
Tasa de acierto	**71,3%**	70,9%	68,2%
Tasa de error	**28,7%**	29,1%	31,8%
Recall mat.	**76,7%**	76,1%	73,1%
Recall no mat.	65%	**68,5%**	66,1%
Fp rate	**35%**	35,3%	38,2%
Precision mat.	**72,1%**	71,7%	69,7%
Precision no mat	**70,2%**	67,1%	66,1%
F-Measure	**74,3%**	73,8%	71,4%
Estimación matrícula	3.243	3.046	3.184

Fuente: Elaboración propia

En la tabla de comparación para la base de validación se observa que el modelo logit base es el modelo que cuenta con los mejores estimadores de los tres modelos, salvo en el Racall de los no matriculados, que el mejor estimador pertenece al modelo logit con interacciones, y en el fp rate donde es mejor el árbol de decisión, para todos los demás el modelo logit base es el mejor, al igual como ocurría para la base de calibración. En segundo lugar queda el modelo logit con interacciones, ya que tiene mejores estimadores que el árbol de decisión, que queda en el tercer lugar, igual a lo que se obtenía para la base de calibración.

Al comparar los dos modelos logit se tiene que el modelo logit base es más preciso en la estimación que el logit con interacciones, ya que cuenta con los mejores estadísticos, a nivel general, solo es menor en el Recall de los no matriculados, un 3,5% más bajo, pero en todos los demás el modelo es el mejor.

Además, para realizar una comparación adicional entre los modelos logit se revisarán la log-verosimilitud, y los criterios de información AIC y BIC, que se obtuvieron para las bases de calibración.

Tabla 43: Comparación modelos logit.

Estadístico	Logit base	Logit con interacciones	Diferencia
Log-verosimilitud	-12594.8	-12725	130,2
AIC	25265,6	25534	268,4
BIC	25196,9	25457,4	260,5
Ratio verosimilitud	0,19	0,19	0

Fuente: Elaboración propia.

Se puede observar que los valores para los estadísticos son similares, no existe mucha diferencia entre ambos modelos, aunque son levemente mejores para el modelo logit base, ya que se busca que los estimadores sean lo más bajo posible, por lo que en este caso, al comparar según estos estadísticos el modelo logit base es mejor que el modelo con interacciones.

Luego de comparar los tres modelos desarrollados anteriormente se puede concluir que el mejor modelo, medido de acuerdo a los indicadores, en cuanto al ajuste a los datos y a la validación de los mismos viene siendo el modelo logit base. Esto no quiere decir que los otros dos sean malos, solo que este modelo presente mejores estimadores de clasificación.

4.6.1.2 COMPARACIÓN MODELOS LOGIT CON MODELO DE LA EMPRESA

Para realizar la comparación entre los modelos logit aquí desarrollados y con el modelo que cuenta la empresa se utilizará como datos de validación la base que cuenta con información del año 2015.

La comparación se realizará en tres ítems:

- Los estadísticos de la tabla de confusión
- Cantidad estimada de alumnos que se matriculan, y
- Error de conversión

Tabla 44: Comparación base 2015.

Estadístico	Logit base	Logit con interacciones	Logit Penta
Tasa de acierto	**70,2%**	69,8%	69,8%
Tasa de error	**29,8%**	30,2%	30,2%
Recall mat.	**74,3%**	70,7%	70%
Recall no mat.	65,7%	68,8%	**69,6%**
Fp rate	34,1%	31,1%	**30,3%**
Precision mat.	71,1%	72%	**72,3%**
Precision no mat	**69,3%**	67,4%	67,1%
F-Measure	**72,6%**	71,3%	71,1%

Fuente: Elaboración propia.

Al realizar las comparaciones de los modelos para estos datos de validación se observa que los estadísticos son muy similares, en caso de que un estadístico sea mejor que el de otro modelo es por poco, así que todos ganan en un aspecto. Sin embargo, el modelo logit base es superior en cinco de ocho estadísticos, entre ellos la tasa global de aciertos, es decir, en promedio tiene una mejor estimación a nivel general. Otro estadístico en el que es superior es en Recall para matriculados, es decir, a quienes el modelo estima que se matricula y realmente lo hacen. Por otra parte el modelo con interacciones es mejor en tres estadísticos que el modelo de la empresa, igualan en dos y es peor en tres estadísticos.

Al comparar el modelo entre el modelo logit base con el modelo con interacciones se obtiene que el logit base es mejor en cinco de ocho, misma cantidad que para el modelo de la empresa sponsor.

Por estos motivos, si se tuviera que elegir un modelo en base a estos criterios el mejor sería el logit base, ya que cuenta con una mayor cantidad de estadísticos que son levemente superiores en comparación a los otros. El segundo y tercer aspecto para la comparación de los modelos, cantidad estimada de matrícula y error promedio de conversión, se presenta en la siguiente tabla.

Tabla 45: Comparación de modelos.

Detalle	Logit base	Logit con interacciones	Penta	Valor real
Pronóstico	7343	**6902**	6794	7022
Error de pronóstico	4,5%	**-1,7%**	-3,2%	
Probabilidad promedio modelo	53,0%	51,0%	49,9%	~53,1%
# matriculados promedio del modelo	7008	6742	6552	
Error promedio modelo	**-0,2%**	-3,9%	-6,7%	
Error promedio conversión	**5,7%**	6,6%	6,9%	
Tasa de aciertos modelos	**59%**	52%		

Fuente: Elaboración propia.

Se puede apreciar que el modelo logit base sobreestima la cantidad de matriculados, predice que se matriculan 7343 cuando en realidad se matricularon 7022, por lo que comete un error de un 4,5% de sobrestimación. Por otra parte el modelo con interacciones subestima, la cantidad de matriculados, ya que predice 6902, un 1,7% bajo el valor real. Al igual que el modelo con interacciones el modelo de la empresa subestima la cantidad de matriculados en un 3,2%.

Obteniendo la probabilidad promedio del modelo, es decir, la suma de las probabilidades de cada postulante dividido por el total de ellos se obtiene que el modelo que se acerca más a la probabilidad real de matrícula es el logit base, pues se equivoca en un 0,1% a diferencia del modelo con interacciones que tiene una probabilidad 2% menor y el modelo de la empresa que es 3,2% más bajo. Con estas probabilidades promedio se procede a calcular la cantidad estimada promedio de matriculados y el modelo que se encuentra más cerca del valor real es el modelo logit base, con un promedio de 7008 matriculados, lo que da un error de -0,2%, para el modelo con interacciones es de -3,9% y para el modelo de la empresa es de un -6,7%. Por lo que en este sentido el mejor modelo sería el logit base, pues está muy cercano al valor real de matriculados.

El último ítem para realizar la comparación entre los modelos es el error promedio de conversión. Para poder calcular este error se parte tomando la esperanza de los alumnos que postulan a una carrera, con lo que se obtiene la conversión esperada para esta y luego se compara con la conversión real que tuvo el programa, con lo que se obtiene el error por cada una, y luego se obtiene el promedio de los errores de todas las carreras. Dicho esto, se procede a explicar los resultados obtenidos para este ítem.

Observando los datos de la tabla 45 se aprecia que el modelo que tiene menor error promedio de conversión es el modelo logit base con un 5,7% versus un 6,9% que es el error promedio del modelo de la empresa. Aquellas carreras donde se tiene un mayor error es en las cuales sufrieron una variación en la cantidad de convocados, por lo que este es un punto de alerta que se debe tener en cuenta. Junto con esto se tiene un mayor error en aquellas carreras donde una gran cantidad de convocados postulan en primera preferencia pero no se terminan matriculando, de igual forma en aquellas carreras que no son muy demandadas en la primera preferencia, pero se matriculan una gran cantidad de convocados se tiene un error por sobre el promedio. Este error es más difícil de predecir que la variación en la cantidad de los convocados, por esto se podrían usar ambos modelos de forma que se complementen entre sí.

Además comparando la cantidad de carreras que el logit base estima mejor que el modelo de la empresa, utilizado como comparación, es mayor, ya que precide un total de 75 carreras mejor de 127, lo que equivale al 59% de los programas, por lo que el modelo de la empresa estima mejor un total de 52 carreras lo que equivale al 41% del total de programas.

Por otra parte se tiene el modelo logit con interacciones, que tiene un error promedio de 6,6% versus el 6,9% que posee el modelo de comparación. Esto le permite estimar mejor un total de 66 carreras, lo que equivale al 52% del total de programas impartidos por la universidad, versus las 61 carreras que estima mejor el modelo de la empresa lo que corresponde al 48% de la oferta académica. Por lo que si bien el modelo logit con interacciones es mejor que el modelo de la empresa es mejor, esta mejora es marginal, ya que la diferencia no es muy amplia.

Ya se observó que ambos modelos logit son mejores que el modelo creado por la empresa. Por lo que comparando estos modelos, que fueron desarrollados para la presente memoria, se tiene que el logit base es mejor tanto en el error promedio de conversión como en la cantidad total de programas que estima mejor, ya que en el primer aspecto se tiene que el logit base tiene un error de 5,7% versus el 6,6% que posee el con interacciones. De igual forma, el logit base estima mejor un total de 75 programas, 9 por sobre el modelo con interacciones que estima mejor 66.

Finalmente, se pudo apreciar que el modelo logit base es superior en dos de tres ítems de comparación, donde es mejor que los otros dos modelos, y les saca una gran diferencia. Por otra parte se tiene el modelo con interacciones, donde si bien es peor en dos de tres ámbitos en comparación con el logit base le gana en los tres aspectos de medición que el modelo desarrollado por la empresa, aunque este es marginalmente mejor en dos de ellos, ya que la mejora no es muy grande, de igual forma es más preciso. Entonces lo interesante acá es que se crearon dos modelos, y que ambos son buenos dependiendo del indicador que se utilice para medir la exactitud en la predicción del modelo, y que son, en general, mejores que el modelo con el que cuenta actualmente la empresa, además de que se complementan, pues uno es mejor que el otro en ítems distintos, aunque claro, lo ideal hubiese sido que un modelo fuera superior que el que tiene la empresa en todos los aspectos.

4.7 MODELO LOGIT BINOMIAL POR DÍA

Para el desarrollo del modelo por día se utilizará el modelo logit base, pues es el que tiene el menor error promedio en la conversión por carrera, por lo que lo hace más idónea para lo que se busca, que es estimar la conversión de la carrera por día para poder realizar una comparación entre la matrícula real del programa y el estimado y poder generar las acciones oportunas a tiempo, como por ejemplo, si se sobreestimó una carrera con el modelo, poder generar acciones para aumentar la cantidad de alumnos, o por otro lado, si se subestimó mover los recursos que estaban destinados a esa carrera hacía otra que los necesite más. También se podrá realizar una comparación entre la matrícula actual y como se encontraba la carrera el mismo día del año anterior con lo cual se podrán generar alarmas a tiempo y no esperar hasta el final del proceso de admisión para darse cuenta que una carrera estuvo más baja y generar acciones cuando ya sea muy tarde y más costoso. Lo que se busca es ser proactivo y no reactivo durante este período que es de vital importancia para la institución.

4.7.1 DESARROLLO DEL MODELO POR DÍA

Para el desarrollo del modelo se utilizarán las mismas variables que el logit base, ya que es una extensión del mismo, sólo que ahora se hará para los matriculados hasta el día uno, luego con los matriculados hasta el día dos, que incluye los del primer día, y luego para los matriculados hasta el tercer día, que vendría siendo el mismo modelo logit base que ya fue realizado. Por lo tanto las variables que se ingresarán al modelo serán las mismas, y que listan a continuación.

Tabla 46: Lista de variables ingresadas al modelo.

Información postulante	Interacción postulante-universidad	Información de Mercado
Colegio de procedencia	Orden de postulación	Conversión anterior
Quintil	Simuló misma carrera	Variación carrera en el mercado
Año egreso	Log # simulaciones	Carreras con baja conversión
Puntaje PSU sobre el promedio de la carrera	Certificado	Área de conocimiento
Beca externa	Beca simulada	Valor promedio matrícula IP
CAE	Difusión	Conversión anterior
Nivel de estudios del padre	Log # exposiciones a difusión	Variación carrera en el mercado
Postula a la misma comuna de la sede	Sede	Carreras con baja conversión
	Diferencia beca simulada y real	
	Variación beca simulada	

Fuente: Elaboración propia.

Al igual que en el modelo logit base, las variables se dividen en tres grandes grupos:

- Información del postulantes
- Interacción entre postulante y la universidad
- Información de mercado

4.7.2 RESULTADOS DEL MODELO

Los resultados del modelo se pueden observar en la tabla 47. En ella se presenta la variable utilizada, el valor del coeficiente y la significancia del mismo, ya que se da el caso de que una variable es significativa para el tercer día, pero no así para el segundo o el primero. Esto con el objetivo de poder hacer comparable los modelos y ver cómo afecta la variable a los postulantes según el día de matrícula en que se encuentre.

Las variables están ordenada según el nivel de importancia que tienen estas en el tercer día de matrícula.

Se observa que para el primer día se tiene dos coeficientes que son mayores a 1, y la variable más importante en este día es si el alumno postula en primera preferencia. Situación que cambia durante los siguientes días, ya que si bien esta variable sigue siendo importante, no es la que más afecta en la probabilidad de postulante. Luego

le sigue la conversión anterior de la carrera como segunda variable con mayor importancia para el primer día. Esto hace sentido, ya que si una persona postula en primera preferencia quiere decir que esa es la opción que la persona más desea, por lo que este día las personas se matriculan por la marca que representa la universidad.

En cambio para el segundo día la variable más importante es la conversión anterior de la carrera seguida por la postulación en primera preferencia. Este día se asemeja más al tercer día, ya que las primeras cuatro variables en nivel de importancia son las mismas, solo varía un poco en el orden de ellas. Para el segundo y tercer día se puede observar que la conversión de la carrera es la variable más importante, por lo que a diferencia del primer día que importa la universidad, acá tiene una mayor importancia la carrera, si le fue bien le debería ir bien nuevamente.

Se puede observar que para el primer día, a diferencia de los otros dos, los coeficientes de las variables que contienen la información de los beneficios económicos *beca externa, CAE, beca interna simulada,* son menores, por lo tanto se puede asumir que para el primer día son menos importantes. Los postulantes que se matriculan en este día son menos sensibles a estos beneficios en comparación a los otros dos días. Por lo que se puede asumir, que durante el primer día importa más la preferencia por la universidad, es decir, la marca que ella posea por sobre el beneficio económico que puede entregar al postulante para matricularse ahí. Esto también se puede observar en el valor de los coeficientes de las variables *variación entre beca simulada y beca real* y *variación promedio beca simulada* donde sus valores son menores que para el segundo y tercer día. De hecho la última variable ni siquiera es significativa el primer día. Por lo que el primer día es importante la marca y prestigio de la universidad.

Tabla 47: Coeficientes y significancia modelo por día.

Variable	Día 1		Día 2		Día 3	
	Coeficiente	P-valor	Coeficiente	P-valor	Coeficiente	P-valor
Conversión anterior	1,02	0,00	1,73	0,00	2,14	0,00
Ptje sobre ptje promedio de la carrera	0,63	0,00	1,07	0,00	1,43	0,00
Postula 1ra preferencia	1,29	0,00	1,36	0,00	1,32	0,00
Variación entre beca simulada y beca real	0,74	0,02	1,20	0,00	1,28	0,00
Colegio Particular	0,64	0,00	0,79	0,00	0,77	0,00
Variación promedio beca simulada	0,05	0,75	0,33	0,02	0,62	0,00
Egresó un año antes del proceso	0,52	0,00	0,67	0,00	0,60	0,00
Log # de simulaciones	0,54	0,00	0,68	0,00	0,59	0,00
Postula 2da preferencia	0,52	0,00	0,52	0,00	0,55	0,00
Emité certificado	0,54	0,00	0,53	0,00	0,50	0,00
Variación carrera en el mercado	0,40	0,00	0,54	0,00	0,49	0,00
Egresó dos años antes del proceso	0,26	0,01	0,41	0,00	0,48	0,00
Beca externa	0,24	0,00	0,41	0,00	0,45	0,00
CAE	0,28	0,00	0,36	0,00	0,41	0,00
Egresó tres años antes del proceso	-0,03	0,81	0,23	0,03	0,38	0,00
Simula misma carrera que postula	0,35	0,00	0,32	0,00	0,33	0,00
Postula 3ra preferencia	0,35	0,00	0,33	0,00	0,31	0,00
Egresó mismo año del proceso	0,38	0,00	0,42	0,00	0,24	0,00
Log beca interna simulada	0,13	0,00	0,17	0,00	0,23	0,00
Nivel estudios padre universitario	0,15	0,00	0,11	0,02	0,19	0,00
Colegio Subvencionado	0,23	0,00	0,23	0,00	0,19	0,00
Postula a una sede de su misma comuna	0,22	0,00	0,22	0,00	0,19	0,00
Nivel estudios padre CFT-IP	0,18	0,00	0,18	0,00	0,17	0,00
Nivel estudios padre ed. Media	0,16	0,00	0,11	0,01	0,12	0,00
Carreras del área salud	0,19	0,00	0,18	0,00	0,12	0,00
Difusión	0,11	0,01	0,08	0,03	0,11	0,01
Sede República	0,02	0,70	0,00	0,90	-0,08	0,05
Sede Casona	-0,06	0,25	-0,11	0,01	-0,22	0,00
Sede Bellavista	-0,05	0,48	-0,08	0,25	-0,24	0,00
Quintil 4	-0,23	0,00	-0,30	0,00	-0,30	0,00
Año 2015	-0,12	0,14	-0,32	0,00	-0,41	0,00
Sede Los Leones	-0,14	0,22	-0,24	0,02	-0,41	0,00
Quintil 3	-0,48	0,00	-0,54	0,00	-0,52	0,00
Quintil 2	-0,63	0,00	-0,76	0,00	-0,74	0,00
Quintil 1	-0,86	0,00	-0,92	0,00	-0,85	0,00
Carrera con baja conversión y valor matrícula promedio IP	-0,96	0,00	-1,12	0,00	-1,26	0,00
Constante	-4,75	0,00	-4,81	0,00	-4,59	0,00

Fuente: Elaboración propia.

También se puede observar que la variable *puntaje ponderado sobre el promedio* no es muy importante para el primer día como para los otros dos. De igual forma *colegio particular*, que si bien es importante, no lo es tanto como para los otros días.

Otra variable que es interesante observar es la *variación de la carrera en el mercado* donde el valor del coeficiente es menor para el primer día y más alto para el segundo, por sobre el tercero, esto puede indicar que el primer día no importa tanto como le ha ido a la carrera, sino que el alumno se matricula porque eso es lo que él quiere, en cambio el segundo día podría interpretarse como al alumno que no está muy seguro y opta una carrera que sea cotizada por el mercado, más que porque sea de su total preferencia.

Caso aparte son las carreras del área de la salud, ya que esta variable tiene una mayor importancia para el primer día que para los demás, teniendo un coeficiente muy similar para el segundo día y este decae para el tercero. Esto indica que estas carreras si son demandas el primer día, podría ser que el alumno que postula a una de estas carreras quiere matricularse pronto para asegurar su cupo (siendo que si el alumno es convocado tiene su cupo asegurado hasta el tercer día), pero el postulante puede sentir que si no se matricula pronto lo puede perder, ya que estas carreras son muy demandadas.

La variable que contiene la información sobre las simulaciones del postulante se van haciendo más importantes a medida que van transcurriendo los días, lo que tiene correlación con la beca interna, que también va aumentando a medida que avanzan los días. Porque el primer día el beneficio económico es menos importante, aunque la variable certificados es más importante para este día, por lo que quienes lo emiten lo van a hacer efectivo.

También se puede apreciar que las variables *sedes* no son tan negativas para el primer día, incluso no son significativas al 90%, por lo que reafirma que durante el primer día la marca de la universidad es más importante.

La variable que contiene información sobre el nivel de ingreso también es menos negativa para el primer día que para los otros dos, por lo que se podría decir que las personas que se matriculan durante este día son menos sensibles al precio que el resto de los postulantes. A excepción del *Quintil 1* que su coeficiente es 0,01 mayor que para el tercer día, pero menor que para el segundo. Esto podría ser debido a que después la beca se hace más importante en la decisión de los postulantes.

Finalmente, la variable que contiene información sobre el valor promedio de la matrícula de los institutos profesionales también es menos negativa para el primer día que para los otros dos, se podría explicar por la sensibilidad al precio de los postulantes que se matriculan durante este día.

4.7.3 CALIDAD DE AJUSTE DEL MODELO

Para medir la calidad de ajuste del modelo se presentarán las tablas de confusión para cada uno los días, a excepción del tercero que ya fue exhibida, y luego la tabla con los estadísticos para poder compararlos.

Tabla 48: Tabla de confusión primer día.

Observado		**Pronosticado**		
		matricula		Porcentaje correcto
		Si	No	
matricula	No	15.555	1.060	93,6%
	SI	4.505	1.297	22,4%
Porcentaje global				75,2%

Fuente: Elaboración propia

Tabla 49: Tabla de confusión al segundo día.

Observado		**Pronosticado**		
		matricula		Porcentaje correcto
		0	1	
matricula	0	9.307	3.104	75,0%
	1	3.379	6.627	66,2%
Porcentaje global				71,1%

Fuente: Elaboración propia

Tabla 50: Estadísticos modelos por día.

Estadístico	1er día	2do día	3er día
Tasa de acierto	75,2%	71,1%	72,6%
Tasa de error	24,8%	28,9%	27,4%
Recall mat.	22,4%	66,2%	77,6%
Recall no mat.	93,6%	75%	66,5%
Fp rate	6,3%	25%	33,4%
Precision mat.	55%	68,1%	74,1%
Precision no mat	77,5%	73,3%	70,6%
F-Measure	31,9%	69,5%	75,8%

Fuente: Elaboración propia

Como se puede observar de la tabla 50, el modelo para el primer día no tiene una estimación individual muy precisa, su capacidad para predecir a quienes se matricularon es del 22,4%. Esto es debido a que se cuenta con menos datos de matriculados para ese día y van aumentando a medida que avanzan los días, por eso para el segundo día se tienen estadísticos mejores, similares a los del tercer día. Por lo que estos no son buenos estadísticos para medir la precisión del modelo, ya que lo que se busca es tener una estimación de la conversión por carrera para tener un

punto de comparación y poder hacer un seguimiento de las carreras al finalizar el día.

4.7.4 VALIDACIÓN DEL MODELO

Para la validación del modelo se estimará la conversión por día para las carreras y se comparará con la conversión real del año 2015, además se estimará la cantidad de matriculados que predice el modelo a nivel individual para tener un punto de comparación.

Tabla 51: Comparación conversión real y estimada.

Detalle	Día 1	Día 2	Día 3
Matrícula real	**3454**	**5857**	**7022**
Pronóstico	1646	5605	7343
Error de pronóstico	-52,3%	-4,3%	4,5%
Probabilidad promedio modelo	26,5%	44,2%	53,0%
# matriculados promedio del modelo	3502	5841	7008
Error promedio modelo	1,3%	-0,21%	-0,19%
Error promedio conversión	5,3%	5,7%	5,7%

Fuente: Elaboración propia.

Se puede apreciar de la tabla 51 que el modelo para el primer día tiene un error muy grande en la estimación individual de matriculados, pues subestima en demasía la cantidad de postulantes que finalmente se matriculan, llegando a más del 50% de error. Para el modelo del segundo y tercer día se tiene un error similar, cercano al 4,5%, pero con signo contrario, ya que el modelo del segundo día subestima la matrícula individual de postulantes y el modelo del tercer día, como ya se comentó anteriormente, sobreestima la cantidad final de matriculados en una proporción similar, en valor absoluto.

Por otra parte se tiene que tomando la probabilidad promedio del modelo, es decir, sumando la probabilidad de cada postulante y dividiendo por el total de postulantes (13216) se obtiene la probabilidad promedio del modelo. En este aspecto el modelo tiene una probabilidad muy similar a la real, esto se puede observar ya que la cantidad promedio de matriculados, que se obtiene de multiplicar la probabilidad promedio del modelo por la cantidad total de postulantes se obtiene un valor muy cercano al real. Se tiene que el mayor error se produce para el primer día y es de un 1,3%, para el modelo del segundo y tercer día el error es menor al 1%, siendo aproximadamente de -0,2%, negativo ya que en ambos se subestima la cantidad total de matriculados.

Finalmente, observando el último ítem de medición se puede apreciar que para los tres días se tiene un error promedio de conversión de 5,3% para el primer día y de 5,7% para el segundo y tercer día. Por lo que para el primer día el modelo ajusta mejor la conversión que para los otros dos días, pero la diferencia no es muy grande.

El modelo presenta un alto error en la estimación individual de matrícula para el primer día, donde el error supera el 50%. Sin embargo, para el día dos y tres tiene un error cercano al 4,5%, lo cual se encuentra en un rango aceptable, pero es mucho mejor que para el día uno. Por otra parte, la probabilidad promedio del modelo se acerca mucho a la probabilidad real, por lo que al obtener la matrícula promedio del modelo esta se acerca bastante a la real, donde el error del primer día es de 1,3%, mucho más bajo que para el error del pronóstico, y para el segundo y tercer día es del 0,2%, con lo que la cantidad promedio de matriculados es muy similar a la real. En cuanto a la conversión, que es lo que se busca con este modelo por día, el error promedio varía entre 5,3% y 5,7%, con lo que se tiene un buen modelo para predecir la conversión de las carreras por día.

Con la creación de este modelo se tiene un punto de comparación entre las carreras por día para el estimado y para el nivel real, con lo que si el primer día la carrera está bajo el estimado se pueden generar acciones a partir del segundo día, con lo que no se debe esperar al tercer día para darse cuenta de este aspecto. Más aún, si el modelo pronostica una carrera baja se pueden idear acciones desde el primer día y se obtendría un día de ganancia. Con esto se permite generar alertas sobre las carreras que se pronostiquen bajas e ir midiendo su evolución durante el día de manera más detallada para ver si se cumple o no lo que dice el modelo y poder actuar a tiempo.

Este modelo también permite tener una comparación día a día con el año anterior y así tener otro punto de comparación y poder generar alarmas en aquellas carreras que estén por debajo de lo esperado, ya sea con respecto al modelo o en comparación al año anterior y generar campañas para aumentar la matrícula. Permite hacer un seguimiento con menor detalle las carreras que se pronostiquen con buena conversión durante el día y que en comparación al año anterior vaya en un nivel similar o superior, y destinar todos los esfuerzos en mejorar la conversión de aquellas carreras bajas.

5 CONCLUSIONES

La memoria se desarrolla en el contexto del proceso de admisión universitario para una universidad que se encuentra adscrita al sistema único de admisión, más específicamente para el primer período de matrícula del proceso, que tiene una duración de tres días. Para el desarrollo del trabajo se cuenta con información que le proporciona el DEMRE a la casa de estudio, que contiene información socioeconómica del alumno, los puntajes obtenidos en cada una de las pruebas y las postulaciones que este realizó en el proceso, además se contó con información que la universidad entrega a la empresa Penta Analytics y que tiene relación con la interacción que tuvo el alumno con ella.

Con la información disponible se procede a resolver el objetivo general de la presente memoria que es estimar la probabilidad de matrícula de un postulante convocado por la universidad durante los tres primeros días de matrícula DEMRE, a través de un modelo de elección discreta con variables explicativas. Para luego poder estimar la conversión que tendrán las distintas carreras que ofrece la universidad y poder generar acciones desde el primer día en aquellas carreras donde la conversión se pronostique baja.

Para cumplir el objetivo general se desarrollaron tres modelos, todos a nivel individual. Dos de ellos fueron modelos de elección discreta, Logit Binomial, y un modelo de clasificación, Árbol de Decisión, con el objetivo de determinar cuál de ellos era el que ajustaba y predecía mejor los datos con los que se contaba.

Los principales resultados obtenidos de los modelos logit son, que las variables que más afectan sobre la probabilidad de matrícula de los postulantes son en primer lugar la conversión anterior que tuvo la carrera a la cual postularon, es decir, que si postulan a una carrera que fue demanda el año pasado aumenta la probabilidad de que se matriculen. Le sigue el puntaje ponderado del alumno por sobre el puntaje promedio de la carrera, si tuvo un puntaje alto con respecto al promedio de los convocados es más probable que se matricule. En tercer lugar para el modelo logit base se encuentra la postulación en primera preferencia, en cambio para el logit con interacciones se encuentra la variación entre la beca real y simulada por el postulante. La interpretación de ambas variables es la misma para los modelos, solo cambia el orden de importancia en cada uno.

Los principales resultados del modelo logit base son:

- Si un alumno postula en primera preferencia a la universidad, la probabilidad de matrícula aumenta en promedio un 15,6% por sobre un alumno de las mismas características que postula en la segunda preferencia. Esta diferencia aumenta a un 21% entre postular en primera versus tercera preferencia, llegando a un 27% aproximadamente entre un postulante que lo hace en primera versus una opción posterior a la tercera. Estas diferencias son

similares para ambos modelos Logit. Con lo cual se valida la hipótesis de que el orden de postulación afecta la probabilidad de matrícula.

- Si un alumno simula la misma carrera a la que postula la probabilidad de matrícula aumenta en promedio un 6,6% por sobre alguien de las mismas características que no lo hace. Una diferencia mayor se encuentra entre quienes simularon, pero no emitieron el certificado, pues aquellos postulantes que lo emitieron al momento de simular una beca su probabilidad aumenta un 10% aproximadamente frente al mismo alumno que no emitió el certificado. La hipótesis de que la interacción entre universidad y postulante afecta la probabilidad de matrícula es validada.

- Si los postulantes cuentan con beneficios económicos externos, otorgados por el Estado, como la beca externa y el CAE, su probabilidad de matrícula es un 8% más alta aproximadamente que si no contaran con estos beneficios. Por lo que la hipótesis de que los beneficios económicos aumentan la probabilidad de matrícula de los alumnos se valida.

En cuanto al modelo logit con interacciones los principales resultados obtenidos de las interacciones entre variables son los siguientes:

- En relación a la interacción entre tipo de colegio y beca de arancel interna se obtiene que, un alumno que cuenta con beca y proviene de colegio particular tiene una probabilidad de matrícula un 10,1% más alta que si el alumno fuese de colegio subvencionado y que cuente con beca. Y esta diferencia de probabilidad es aún mayor para un estudiante que provenga de colegio particular, pero que no cuente con beca, esta diferencia asciende a un 14,3%. Esta brecha es mucho menor para un postulante que provenga de colegio subvencionado y que cuente con beca y el mismo alumno que no tenga beca, la diferencia es de un 5% aproximadamente, por lo que el postulante de colegio particular es mucho más sensible a la beca interna que el de colegio subvencionado por lo que se debería becar este segmento para aumentar la conversión, ya que su probabilidad de matrícula es más alta frente a postulantes de otros colegios.

- Situación similar ocurre con postulantes de colegios particulares y actividades de difusión, pues si el alumno estuvo expuesto a actividades de difusión y proviene de este tipo de colegio tiene una probabilidad de matrícula un 11% más alta que si no estuvo expuesto a difusión.

- Finalmente, quienes postulan a una misma sede pero pertenecen a diferentes quintiles de ingreso tienen una probabilidad de matrícula distinta. Si un

alumno que pertenece al quintil 2 postula a la sede Los Leones su probabilidad de matrícula disminuye un 5,2% frente a que postulara a la sede de Casona y disminuye en un 10,5% si postula a la misma sede en vez de que lo hiciera a la sede de República, y si postula a la sede Los Leones en vez de postular a las sedes de Concepción, Viña o Bellavista su probabilidad disminuye en un 13%. De igual forma, si pertenecen a distintos quintiles pero postulan a la misma sede también se observa que la probabilidad de matrícula es distinta. Si una persona que postula a la sede de Casona y pertenece al primer quintil de ingreso tiene una probabilidad un 3,1% menor que si perteneciera al quintil 2 y si en vez de pertenecer al quintil 1 pertenece al quintil 3 esta diferencia de probabilidad aumenta a un 10,1%. Por lo tanto se encuentran diferencias entre sedes y quintiles, a mayor ingreso socioeconómico se tiene una mayor probabilidad de matrícula.

Para el árbol de decisión, las variables que son importantes para predecir la matrícula del los postulantes son si simula en la misma carrera, si postula en primera preferencia, si emite certificado, y si tiene un puntaje ponderado sobre el promedio mayor a 0,4. Esta última variable se encuentra estandarizada, por lo que un puntaje ponderado sobre el promedio de 0,4 significa que la diferencia entre el puntaje del postulante y del promedio de la carrera no sea inferior a 90 puntos. El 86% de las personas que pertenecen a este grupo se matriculan en la universidad.

En la comparación de los modelos, para la base de calibración, para los estadísticos de clasificación, como Recall, Precision, F-Measure y porcentaje global de acierto, los que tuvieron un mejor desempeño fueron los modelos Logit por sobre el árbol de decisión. El modelo logit base tuvo un Recall de 77,6%, el modelo logit con interacciones un Recall de 73,2% y el árbol un 72,6%, una diferencia del 5% entre el mejor modelo y el peor, en cuanto al estadístico Precision de los modelos, nuevamente el logit base es el mejor modelo con un porcentaje del 75,8%, el logit con interacciones de un 73,9% y en último lugar el árbol con un 71,4%. Finalmente, para la tasa global de aciertos, el logit base continúa siendo el mejor modelo con un 72,6% de aciertos totales versus un 71,2% para el logit con interacciones y un 69,5% para el árbol de decisión, que si bien en este aspecto acorta las diferencias sigue por debajo de los otros dos modelos. En cuento a las ventajas del árbol de decisión sobre el modelo logit se encuentra la facilidad en la interpretación del mismo, ya que es un modelo simple y visualmente fácil de entender, por otra parte el modelo logit permite determinar el efecto individual de cada variable ingresada al modelo en la probabilidad de matrícula. Por lo que analizando los estadísticos y considerando los pro y contras de cada modelo se eligieron los modelos logit para realizar la estimación de la conversión y la comparación con el modelo actual de la empresa.

En lo referente a la comparación con el modelo que cuenta actualmente la empresa se obtuvieron resultados satisfactorios, ya que ambos modelos logit son mejores que el actual según el indicador con el que se mida. Se tiene que el modelo logit con interacciones es el que mejor estima la cantidad individual de matriculados, pues es el que tiene el menor error, un 1,7% versus un 4,5% del modelo logit base y un 3,2%

del modelo de la empresa. Por otra parte el modelo logit base es el que tiene el menor error promedio en la estimación de la conversión, con un 5,7% frente a un 6,6% del logit con interacciones y un 6,9% del de la empresa. Por lo tanto, este modelo, logit base, es el mejor para predecir la conversión de las carreras que ofrece la universidad, ya que posee el menor error de los modelos.
Para la generación del modelo por día se utilizó el modelo logit base, ya que cuenta con el menor error promedio en la conversión de carreras, por lo que, para la finalidad de este modelo es el que tiene una mejor predicción.

En este modelo se encontraron resultados interesantes, como por ejemplo, que para el primer día las variables de beneficios económicos son menos importantes que para el resto de los días, y la variable más importante es si postula en primera preferencia, por esto se puede pensar que para el primer día lo más importante es la marca que la universidad logra crear en los postulantes, por sobre los beneficios económicos que pueda otorgar, por esto la primera preferencia es la variable que más influye en la probabilidad de matrícula donde se demuestra el interés por pertenecer a esa casa de estudio. También se encontró evidencia de que las carreras del área de la salud son más importantes en el primer día que en los demás, esto se puede deber a que son carreras más demandas que las otras que ofrece la universidad, y los postulantes sienten que su cupo se lo pueden quitar (lo cual no es posible) y se van a matricular durante el primer día. El segundo y tercer día tiene un comportamiento similar donde importa más la carrera, conversión anterior, y las becas tienen una mayor importancia que el día 1, esto puede deberse a que para estos días los postulantes están más pendientes de que sean carreras “buenas” donde obtengan una mayor ayuda económica para poder estudiar.

En cuanto a la conversión estimada por carrera por día, se obtuvieron resultados satisfactorios. Para el primer día se tuvo un error promedio de un 5,2%, menor incluso que para el día dos y tres donde fue de 5,7%.

Finalmente, se concluye que el trabajo realizado fue exitoso, logrando en primer lugar el objetivo general y cada uno de los objetivos específicos. Se crearon dos modelos que son mejores que con el que cuenta actualmente la empresa según el indicador con el que se mida la mejora. También se encontraron resultados relevantes en cada uno de los modelos realizados. Además, el trabajo permite entender de mejor forma el problema que se quiere resolver, ya que se debe resolver constantemente para actualizar el comportamiento de los postulantes, por lo que este trabajo es muy útil. Junto a esto permite identificar qué variables y acciones de marketing que puede manejar la universidad influyen más en la matrícula, y también cuanto influye cada, con lo que se puede identificar mejor las acciones más efectivas y colocar más esfuerzo en ellas, como lo son las simulaciones y la emisión de certificados por sobre la difusión que se realiza. Esto da luces de que lo que se hace actualmente tiene un amplio margen de mejora para hacerla más efectiva, como se puede apreciar en la tabla 22, donde la variable relativa a preunab no es significativa, donde se destina aproximadamente un cuarto del presupuesto de actividades de Difusión, alrededor de los $230 millones de pesos, por lo que este trabajo ayuda a la universidad a darse cuenta que estas acciones no están teniendo el efecto esperado y

así poder rediseñarlas, cambiar el foco o realizar una reasignación del presupuesto. Por otra parte, este trabajo también es una primera aproximación para poder realizar una asignación de becas durante el período de matrícula, ya que se cuenta con un presupuesto acotado y se debe decidir a qué postulantes becar y a cuales dejar fuera para, en primer lugar, aumentar la conversión y en segundo lugar mantenerse dentro del presupuesto, por lo que esto podría ser visto como una aproximación al "problema de la mochila" de optimización.

6 RECOMENDACIONES

Del trabajo realizado en la presente memoria, sus resultados, análisis y conclusiones se desprenden las siguientes recomendaciones.

En primer lugar se deben potenciar las actividades de difusión. En el análisis descriptivo se mostró como la cantidad de alumnos que estuvieron expuestos a estas actividades ha disminuido en el tiempo, tomando en consideración los años entre 2013 y 2015, tanto en la cantidad total de alumnos convocados como en la proporción que representan dentro los matriculados, cada año disminuye más los alumnos expuestos a difusión que se matriculan. Esto porque en los modelos se encontró evidencia de que éste es un factor que aumenta la probabilidad de matrícula, pero no está siendo muy efectivo en los postulantes. Además, se deberían potenciar estas actividades para alumnos de colegios particulares, ya que en estos aumenta en mayor medida la probabilidad de matrícula.

Situación similar ocurre con las simulaciones, deben fomentar el uso de esta herramienta para las becas internas y la emisión del certificado que valida su beneficio, ya que con este último aumenta la probabilidad de matrícula aún más entre quienes simularon. Y deben hacer énfasis en que los alumnos realicen esta acción en carreras que postulan a la universidad, se encontró que un 7% de quienes simulan no lo hicieron en la misma carrera en la que fue convocado, con lo que disminuye su probabilidad de matrícula pues no conoce los beneficios a los que podrían acceder. Una acción a realizar en este aspecto es que recertifiquen a los postulantes que emitieron su certificado, pero finalmente no fue validado (un certificado queda inválido cuando se ingresan datos que posteriormente no son los finales, por ejemplo que genera un certificado con un puntaje PSU incorrecto o que es generado suponiendo que cuenta con beca externar o CAE y finalmente no le otorgan estos beneficios), ya que el postulante con certificado válido tiene una probabilidad un 10% más alta que si no lo tuviera.

En cuanto a las postulaciones, debe potenciar la primera preferencia, era una de las variables que más influía en la probabilidad de matrícula. Podrían fortalecer esto otorgando beneficios adicionales a los que actualmente entregan, claro que se tendría que hacer un estudio sobre esta medida, pero con esto podrían hacer más atractiva la opción de postular en primera preferencia a la universidad. Junto a esto, si bien postular en primera opción es importante, no deberían cortar las

postulaciones en la sexta preferencia, ya que el año 2014 aproximadamente 300 personas postularon bajo la sexta, de las cuales 60 se matricularon en la universidad, por lo que al cortar en esta preferencia se podría perder una cantidad similar de alumnos. Considerando que no tiene un mayor costo económico para la universidad y pueden aumentar la cantidad de convocados en estas preferencias, sobre todo teniendo en cuenta la entrada en vigencia de la gratuidad.

En el ámbito de los beneficios internos se encontró evidencia que los alumnos de colegios particulares son más sensibles a esta beca, por lo que este es un resultado que se debe tener en cuenta al momento de la matrícula. Además, el modelo por día realizado mostró que la sensibilidad de los alumnos a los beneficios internos varía según el día de matrícula, por lo que se podría dar un beneficio según el día en que se encuentre. Si bien en este modelo se realizó una primera aproximación, entregó un resultado interesante que puede ser analizado en el futuro. Junto a esto se debe seguir trabajando en potenciar la marca de la universidad para que los beneficios económicos vayan disminuyendo su importancia en el momento de la matrícula.

Es importante realizar un seguimiento a las carreras que se pronostiquen bajas para poder aumentar su matrícula, en caso de que se cumpla esta predicción, realizando acciones a tiempo y siendo proactivo. De igual forma, se deben comparar las carreras que se pronostiquen altas y no dejarlas de lado para que no se vayan a dar sorpresas, de que el modelo predijo alta y finalmente eso no ocurrió. Hay que tener especial atención en aquellas carreras donde la cantidad de convocados fue mucho menor al año anterior, ya que en estos casos el modelo tiende a cometer un error mayor que el promedio y podría ser un indicio de que tendrá baja conversión. Por lo que para tener una estimación más robusta de la conversión por carrera se podrían utilizar ambos modelos en la predicción en busca de disminuir los errores que se puedan cometer en la estimación.

Además de determinar las características que más influyen en la probabilidad de matrícula y poder estimar la conversión de las carreras, se puede aprovechar el modelo para priorizar a los postulantes dentro de la carrera, ya que cada uno tiene una probabilidad distinta dentro de la misma se pueden ejercer acciones sobre aquellos que tengan la probabilidad más alta de matrícula para fidelizarlos y que finalmente se decidan por la universidad y así ir asegurando la conversión de la carrera. También se puede optar por realizar acciones de telemarketing sobre aquellos alumnos que tienen una probabilidad de matrícula media-alta, para asegurar su conversión en la universidad y dejar a los que tienen una esperanza de matrícula muy alta, ya que es más probable que lo terminen haciendo. De esta forma se destinan los esfuerzos hacia aquellos postulantes que se encuentran más indecisos en su decisión y realizar campañas masivas poco focalizadas en aquellos de bajo score, según se estime conveniente.

7 TRABAJO FUTURO

Dentro de los posibles trabajos futuros a realizar, se destaca en primer lugar el medir el efecto que tendrá la entrada de la gratuidad universitaria, que entró a regir este año y será puesta en marcha a partir del proceso de admisión 2016, es decir, el primer proceso de admisión posterior a esta memoria. Por lo que este efecto no pudo ser incluido en el trabajo, ya que no se sabe de qué manera van a reaccionar los estudiantes ni el mercado. Por lo tanto lo primero que se tendría que hacer es ver cómo responden los estudiantes, sobre todo de los cinco primeros deciles de menor ingreso, ante este incentivo para continuar sus estudios. En particular, ver cómo reaccionan ante la universidad que no es parte del conjunto de universidades que puede optar a la gratuidad. Al aventurarse sobre qué efecto podría tener esta condición exógena sobre la matrícula de la universidad se podría suponer que todos aquellos que cuenten con este beneficio o no estén seguros de si lo tienen o no vayan a postular a universidades que tenga gratuidad dentro de sus primeras opciones para poder optar a estudiar sin los costos que esto conlleva, esto podría provocar que se muevan las preferencias dentro de la universidad, es decir, alguien que sin gratuidad hubiese postulado en primera preferencia a la universidad ahora con este beneficio ya lo hace en ese orden, sino que lo hace en la segunda o tercera y como su primera opción elige a una institución que esté adscrita a la gratuidad para poder ocupar el beneficio, ocurriría lo mismo con los postulantes en otras preferencias, esto implicaría que la distribución de las preferencias de los postulantes se debería correr en al menos una.

Otro aspecto de trabajo futuro que se debe tener en cuenta es la actualización del modelo y sus variables introducidas, ya que entre un año y otro pueden variar las características de los alumnos que postulan a la universidad, si bien no se encontraron grandes variaciones en el análisis descriptivo entre un año y otro, es necesario tener actualizado los cambios en las preferencias y sensibilidades los alumnos teniendo en cuenta el primer punto sobre gratuidad, que puede hacer cambiar en demasía las características de los actuales alumnos de la casa de estudio.

En busca de mejorar los modelos y resultados actuales, se puede proponer una nueva metodología, para obtener mejores resultados. Ya que para obtener algunos datos se tuvieron que hacer estimaciones como por ejemplo en la beca interna, que se obtuvo el promedio de las simulaciones que el postulante realizó, también en la obtención de las variables de mercado, lo que pudo restar precisión. Junto a esto también se puede incluir nuevas variables, que entreguen otro tipo de información, más variables de mercado que en este modelo no quedaron consideradas.

De igual forma, se podría realizar un trabajo similar utilizando otro modelo. Uno que podría ser empleado para resolver este problema, podría ser un mixed logit, que viene a mejorar algunas deficiencias del modelo acá usado. Se podría utilizar también un modelo dinámico que emplee la información actual para predecir la matrícula esperada del día siguiente, como una variación del modelo por día.

8 BIBLIOGRAFÍA

[1] Consejo Nacional De Educación, "Conceptos Básicos" [En línea]. Disponible: http://www.cned.cl/public/secciones/SeccionEducacionSuperior/conceptos_basicos.aspx [Fecha de consulta: 29 de Mayo]

[2] Mi Futuro, "Universidades" [En línea]. Disponible: http://www.mifuturo.cl/index.php/donde-y-que-estudiar/universidades [Fecha de consulta: 29 de mayo]

[3] Servicio de Información de Educación Superior, "Informes Anuales: Matrícula" [En línea]. Disponible: http://www.mifuturo.cl/index.php/informes-sies/matriculados [Fecha de consulta: 11 de julio]

[4] J. Barreriro, E. Ruzo, y F. Losada. "Modelo Logit Multinomial: una aplicación regional al sector lacteo." Regional and Sectoral Economic Studies 2004; 4(1):65-86.

[5] DEMRE, "Universidades privadas adscritas al sistema" [En línea]. Disponible: http://psu.demre.cl/proceso-admision/universidades-participantes/universidades-privadas-adscritas [Fecha de consulta: 15 de julio]

[6] Consejo de Rectores de las Universidades Chilenas, "Proceso de Admisión" [En línea]. Disponible: http://sistemadeadmision.consejoderectores.cl/admision_calendario.php [Fecha de consulta: 29 de Mayo]

[7] Universidad Andrés Bello, "Historia" [En línea]. Disponible: http://www.unab.cl/universidad/historia.asp [Fecha de consulta: 29 de Mayo]

[8] La Tercera, "Seis Universidades matricularon menos del 75% de sus seleccionados", [En línea]. Disponible: http://www.latercera.com/noticia/educacion/2013/06/657-527524-9-seis-universidades-matricularon-a-menos-del-75-de-sus-seleccionados.shtml [Fecha de consulta: 29 de Mayo]

[9] A. Mizala, «Determinantes de la elección y deserción en la carrera de pedagogía,» FONIDE, Departamendo de estudios y desarrollo, Santiago, 2011.

[10] Munita Morgan, Juan Pablo. "Contexto social y la elección de ser profesor en Chile." M.S. Tesis, Pontificia Universidad Católica de Chile, Santiago, Chile, 2011.

[11] D. Chapman, «A Model of Student Collegue Choice,» Journal of Higher Education, 1981; 52(5): 490-505.

[12] J. de Dios, y M. Salas. "Modeling educational choices. A binomial logit model applied to the demand for Higher Education." Higher Education 2000; 40(3): 293-311.

[13] F. Barrientos, "Optimización de la oferta de becas para una institución de educación superior", Memoria, Universidad de Chile, Santiago, Chile, 2013.

[14] W. Green, "Econometric Analysis", 5th ed. New York: New York University, Prentice Hall, 2002.

[15] "Logit", notas de clases para IN5602 Marketing II, Departamento Ingeniería Civil Industrial, Universidad de Chile, Primavera 2014.
[16] Karem Padilla, "Identificación de clientes de alto valor para el desarrollo de alianzas de una empresa", Memoria, Universidad de Chile, Santiago, Chile, 2015.
[17] I. Written, E. Frank, M. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", 3rd ed.
[18] M. Vargas, "Rediseño de un modelo de incentivos para vendedores de una tienda por departamentos", Memoria, Universidad de Chile, Santiago, Chile, 2009.
[19] IBM Knowledge Center, "Creación árboles de decisión", [En línea], Disponible: http://www-01.ibm.com/support/knowledgecenter/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/tree/idh_idd_treegui_main.htm?lang=es [Fecha de consulta: 20 de noviembre]
[20] "Variables Explicativas", notas de clases para IN5602 Marketing II, Departamento Ingeniería Civil Industrial, Universidad de Chile, Primavera 2014.
[21] "Evaluación de Modelos", notas de clases para IN5602 Marketing II, Departamento Ingeniería Civil Industrial, Universidad de Chile, Primavera 2014.
[22] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers." Machine learning 2004, 31: 1-38.

9 ANEXOS.

9.1 ANEXO 1: Tramo de Ingreso referenciado en punto 4.1 Selección de variables.

Tabla 52: Tramos ingreso.

Tramo	Desde	Hasta
1	0	144000
2	144001	288000
3	288001	432000
4	432001	576000
5	576001	720000
6	720001	864000
7	864001	1008000
8	1008001	1152000
9	1152001	1296000
10	1296001	1440000
11	1440001	1584000
12	1584001	o más

Fuente: DEMRE.

9.2 ANEXO 2: Conversión por Difusión referenciado en punto 4.4 Análisis Descriptivo.

Tabla 53: Conversión por difusión.

	2015			2014			2013		
	Mat.	Con.	Conver.	Mat.	Con.	Conver.	Mat.	Con.	Conver.
Expuesto a Difusión	3015	5437	55%	3958	6918	57%	4095	7454	55%
No expuesto a Difusión	4007	7779	52%	4253	7888	54%	3125	6453	48%

Fuente: Elaboración propia, datos Penta Analytics.

9.3 ANEXO 3: Conversión por Simulación sobre matriculados referenciado en punto 4.4 Análisis Descriptivo.

Tabla 54: Conversión simulación sobre matriculados.

	Año 2015	Año 2014	Año 2013
Simulan sobre matriculados	81%	84%	77%
Simulan sobre Convocados	43%	47%	40%

Fuente: Elaboración propia, datos Penta Analytics.

9.4 ANEXO 4: Conversión por CAE referenciado en punto 4.4 Análisis Descriptivo.

Tabla 55: Conversión por CAE.

	Año 2015	Año 2014	Año 2013
Matriculados con CAE	5367	6224	5079
Convocados con CAE	9575	10669	9335
Conversión	56%	58%	54%

Fuente: Elaboración propia, datos Penta Analytics.

9.5 ANEXO 5: Proporción matriculados con CAE referenciado en punto 4.4 Análisis Descriptivo.

Tabla 56: Proporción matriculados con CAE.

	Año 2015	Año 2014	Año 2013
Matriculados con CAE	76%	76%	70%

Fuente: Elaboración propia, datos Penta Analytics.

9.6 ANEXO 6: Conversión por Beca Externa referenciado en punto 4.4 Análisis Descriptivo.

Tabla 57: Conversión por Beca Externa.

	Año 2015			Año 2014			Año 2013		
	Mat.	Conv.	Conver.	Mat.	Conv.	Conver.	Mat.	Conv.	Conver.
Beca externa	3773	6490	58%	3377	5372	63%	2650	4598	58%

Fuente: Elaboración propia, datos Penta Analytics.

9.7 ANEXO 7: Proporción matriculados con Beca Externa referenciado en punto 4.4 Análisis Descriptivo.

Tabla 58: Proporción matriculados con beca externa.

	Año 2015	Año 2014	Año 2013
Matriculados con beca externa	54%	41%	37%

Fuente: Elaboración propia, datos Penta Analytics.

9.8 ANEXO 8: Conversión por carrera referenciado en punto 4.4 Análisis Descriptivo.

Tabla 59: Conversión por carrera (1).

Carrera	Sede				Desv. Std.
		2015	2014	2013	
INGENIERIA EN ACUICULTURA	4	0%	88%	43%	0,32
TRABAJO SOCIAL	5	12%	14%	8%	0,04
PSICOPEDAGOGIA	5	22%	27%	31%	0,03
BACHILLERATO EN HUMANIDADES	3	23%	25%	38%	0,09
TRABAJO SOCIAL	3	26%	27%	25%	0,01
KINESIOLOGIA	5	27%	28%	32%	0,03
LICENCIATURA EN FILOSOFIA	1	27%	33%	12%	0,15
TRABAJO SOCIAL	4	28%	46%	25%	0,15
NUTRICION Y DIETETICA	5	31%	31%	32%	0,01
CONTADOR AUDITOR	6	32%	48%	45%	0,02
EDUCACION PARVULARIA	1	33%	45%	27%	0,13
FONOAUDIOLOGIA	1	34%	48%	46%	0,02
DISENNO GRAFICO	4	34%	31%	38%	0,05
ING. EN TELECOMUNICACIONES	3	35%	55%	33%	0,15
PUBLICIDAD	1	35%	40%	38%	0,02
EDUCACION FISICA	5	35%	6%	22%	0,11
LICENCIATURA EN LETRAS	1	36%	41%	35%	0,04
SOCIOLOGIA	3	37%	43%	50%	0,05
CONTADOR AUDITOR	5	37%	37%	54%	0,12
SOCIOLOGIA	4	37%	41%	31%	0,07
BACHILLERATO EN CIENCIAS	5	37%	41%	31%	0,07
PSICOPEDAGOGIA	1	38%	31%	36%	0,04
INGENIERIA EN ADM. EMPRESAS	4	38%	36%	27%	0,06
ODONTOLOGIA	5	39%	50%	44%	0,04
BACHILLERATO EN CIENCIAS	1	39%	55%	50%	0,03
ING SEGURIDAD PREV DE RIESGOS	3	39%	55%	49%	0,04
PERIODISMO	1	40%	50%	53%	0,02
LICENCIATURA EN HISTORIA	4	40%	45%	50%	0,03
TERAPIA OCUPACIONAL	1	40%	44%	44%	0,00
ECOTURISMO	3	40%	52%	49%	0,02
LICENCIATURA EN HISTORIA	1	40%	43%	28%	0,11
DISENNO GRAFICO	1	42%	51%	36%	0,10
NUTRICION Y DIETETICA	3	42%	51%	51%	0,00
INGENIERIA COMERCIAL	5	42%	41%	38%	0,02
LICENCIATURA EN BIOLOGIA	3	42%	69%	66%	0,02
INGENIERIA EN TRANSP MARITIMO	4	43%	92%	89%	0,02

Fuente: Elaboración propia, datos Penta Analytics.

Tabla 60: Conversión por carrera (2).

Carrera	Sede				Desv. Std.
		2015	2014	2013	
ING SEGURIDAD PREV DE RIESGOS	4	43%	51%	43%	0,06
DISENNO DE VESTUARIO Y TEXTIL	1	43%	38%	35%	0,02
PSICOLOGIA	1	44%	43%	49%	0,04
PEDAGOGIA EN INGLES	1	44%	49%	39%	0,07
INGENIERIA EN ADM. EMPRESAS	6	45%	46%	36%	0,07
EDUCACION GRAL BASICA	1	45%	29%	31%	0,02
EDUCACION MUSICAL	1	45%	29%	24%	0,04
ING EN LOGISTICA Y TRANSP	3	45%	60%	50%	0,07
PSICOPEDAGOGIA	4	45%	49%	43%	0,04
ING. CIVIL METALURGIA	5	46%	60%	0%	0,42
FONOAUDIOLOGIA	5	47%	29%	0%	0,21
INGENIERIA EN CONSTRUCCION	3	48%	53%	40%	0,09
INGENIERIA COMERCIAL	2	48%	64%	54%	0,07
KINESIOLOGIA	1	48%	50%	56%	0,04
INGENIERIA CIVIL	5	48%	44%	41%	0,02
ING. CIVIL EN MINAS	4	49%	57%	65%	0,06
LICENCIATURA EN ARTES VISUALES	1	49%	38%	44%	0,04
BIOLOGIA MARINA	3	49%	49%	61%	0,09
DERECHO	2	49%	58%	56%	0,01
TERAPIA OCUPACIONAL	4	50%	61%	57%	0,03
INGENIERIA CIVIL INFORMATICA	4	50%	32%	0%	0,22
NUTRICION Y DIETETICA	4	50%	44%	46%	0,02
EDUCACION FISICA	1	51%	51%	38%	0,09
ARQUITECTURA	4	51%	58%	53%	0,04
DERECHO	5	52%	42%	42%	0,00
TECNOLOGIA MEDICA	5	52%	52%	54%	0,02
ARQUITECTURA	1	53%	50%	47%	0,02
PSICOLOGIA	5	53%	45%	53%	0,06
ING. CIVIL METALURGIA	4	53%	61%	41%	0,14
BACHILLERATO EN CIENCIAS	3	53%	51%	51%	0,00
MEDICINA VETERINARIA	3	54%	54%	48%	0,04
ECOTURISMO	5	54%	49%	42%	0,05
CONTADOR AUDITOR	4	54%	35%	23%	0,09
INGENIERIA EN TURISMO Y HOTEL	4	55%	55%	36%	0,13
INGENIERIA CIVIL INFORMATICA	3	55%	54%	63%	0,06
LICENCIATURA EN FISICA	3	56%	63%	48%	0,10

Fuente: Elaboración propia, datos Penta Analytics.

Tabla 61: Conversión por carrera (3).

Carrera	Sede				Desv. Std.
		2015	2014	2013	
ING AUTOMATIZACION Y ROBOTIC	3	56%	56%	67%	0,08
PEDAGOGIA EN INGLES	5	56%	43%	31%	0,09
ING. ADM. HOTELERA INTERNAC	1	56%	66%	60%	0,04
EDUCACION FISICA	4	56%	58%	41%	0,12
ODONTOLOGIA	3	57%	59%	62%	0,03
BIOQUIMICA	3	57%	60%	56%	0,03
OBSTETRICIA	5	57%	44%	41%	0,02
ING. CIVIL METALURGIA	3	58%	71%	61%	0,07
INGENIERIA CIVIL	3	58%	66%	67%	0,01
TERAPIA OCUPACIONAL	5	58%	61%	47%	0,10
PSICOLOGIA	4	58%	60%	47%	0,09
INGENIERIA EN COMPUT. E INFOR	3	58%	60%	52%	0,05
INGENIERIA EN COMPUT. E INFOR	4	59%	50%	37%	0,09
INGENIERIA EN TURISMO Y HOTEL	1	59%	59%	53%	0,04
KINESIOLOGIA	4	60%	66%	66%	0,00
ING. CIVIL EN MINAS	5	60%	58%	64%	0,04
PEDAGOGIA EN INGLES	4	61%	52%	48%	0,03
INGENIERIA CIVIL INDUSTRIAL	5	62%	44%	39%	0,04
QUIMICA Y FARMACIA	4	63%	58%	66%	0,05
INGENIERIA BIOINFORMATICA	3	63%	45%	22%	0,16
INGENIERIA INDUSTRIAL	3	63%	77%	59%	0,13
DISENNO DE PRODUCTO	1	63%	59%	75%	0,11
ECOTURISMO	4	64%	40%	44%	0,02
INGENIERIA AMBIENTAL	3	64%	65%	62%	0,02
GEOLOGIA	5	64%	73%	61%	0,09
ENFERMERIA	5	64%	66%	65%	0,01
INGENIERIA CIVIL INDUSTRIAL	3	64%	72%	62%	0,07
FONOAUDIOLOGIA	4	65%	70%	65%	0,03
INGENIERIA EN BIOTECNOLOGIA	3	66%	57%	66%	0,06
DERECHO	4	66%	66%	66%	0,00
INGENIERIA COMERCIAL	1	66%	59%	66%	0,05
INGENIERIA COMERCIAL	4	67%	64%	67%	0,01
QUIMICA Y FARMACIA	3	68%	61%	68%	0,05
INGENIERIA GEOLOGICA	3	68%	67%	68%	0,01
LICENCIATURA EN QUIMICA	3	68%	62%	68%	0,05
TECNOLOGIA MEDICA	3	68%	76%	68%	0,05
MEDICINA	3	69%	73%	69%	0,03
ENFERMERIA	3	69%	71%	69%	0,01
TECNOLOGIA MEDICA	4	69%	72%	69%	0,02

Fuente: Elaboración propia, datos Penta Analytics.

Tabla 62: Conversión por carrera (4).

Carrera	Sede				Desv. Std.
		2015	2014	2013	
ING. CIVIL EN MINAS	3	72%	64%	72%	0,05
EDUCACION GRAL BASICA	4	72%	50%	72%	0,16
INGENIERIA CIVIL INDUSTRIAL	4	73%	62%	73%	0,07
INGENIERIA FISICA	3	73%	73%	73%	0,00
EDUCACION PARVULARIA	4	74%	58%	74%	0,11
INGENIERIA EN MARINA MERCANTE	5	75%	78%	75%	0,02
ODONTOLOGIA	4	76%	74%	76%	0,01
BACHILLERATO EN CIENCIAS	4	76%	62%	76%	0,10
INGENIERIA EN MARINA MERCANTE	4	76%	75%	76%	0,01
GEOLOGIA	4	76%	77%	76%	0,00
LICENCIATURA EN ASTRONOMIA	3	77%	79%	77%	0,01
ENFERMERIA	4	77%	79%	77%	0,02
GEOLOGIA	3	77%	78%	77%	0,00
DISENNO DE JUEGOS DIGITALES	1	78%	66%	78%	0,09
INGENIERIA EN BIOTECNOLOGIA	4	80%	84%	80%	0,03
MEDICINA	4	84%	76%	84%	0,05

Fuente: Elaboración propia, datos Penta Analytics.

9.9 ANEXO 9: Conversión real y estimada por modelo logit base referenciado en punto 4.5.1.3 Calidad de Ajuste del modelo.

Tabla 63: Conversión real y estimada por el modelo logit base.

Tramo probabilidad	Conversión Real	Conversión estimada	Error	Cantidad de postulantes
0%-10%	12,0%	7,0%	5,0%	132
11%-20%	14,6%	15,2%	0,6%	876
21%-30%	27,9%	25,1%	2,8%	1791
31%-40%	35,4%	35,0%	0,4%	2416
41%-50%	39,2%	44,8%	5,6%	2955
51%-60%	55,9%	55,1%	0,8%	3864
61%-70%	66,4%	64,9%	1,5%	5376
71%-80%	77,6%	75,1%	2,5%	6408
81%-90%	87,4%	84,8%	2,6%	7677
91%-100%	93,1%	92,2%	0,9%	1770

Fuente Elaboración propia.

9.10 ANEXO 10: Calidad de ajuste del modelo logit con interacciones referenciado en punto 4.5.2.3 Calidad de Ajuste del modelo.

Para medir la calidad de ajuste del modelo se observan los estadísticos definidos anteriormente, comenzando con la tabla de confusión para la base de entrenamiento, calculado con un punto de corte de 0,56.

Tabla 64: Tabla de confusión base entrenamiento.

Observado		**Pronosticado**		
		matricula		Porcentaje correcto
		0	1	
matricula	0	7732	2549	75,2%
	1	4072	8064	66,4%
Porcentaje global				70,5%

Fuente: Elaboración propia.

Además, se obtiene la log-verosimilitud, que nos permite calcular los estimadores como el AIC, BIC y el ratio de verosimilitud.

De la tabla 64 se obtienen los siguientes estadísticos:

Tabla 65: Estadísticos base entrenamiento.

Estadístico	Valor
Log-verosimilitud	-12725
AIC	25534
BIC	25457,4
Ratio verosimilitud	0,19
Tasa de acierto	70,5%
Tasa de error	29,5%
Recall	66,4%
Fp rate	24,8%
Precision	75,9%
F-Measure	70,8%

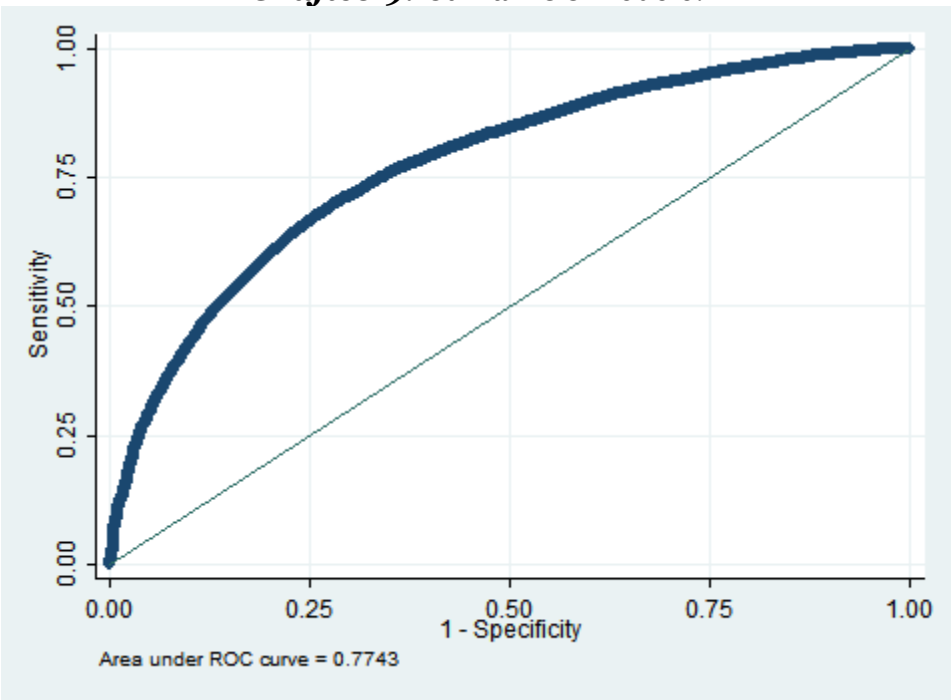
Fuente: Elaboración propia.

Este modelo tiene un AIC y BIC levemente mayores que el modelo a nivel individual sin interacciones y el ratio de verosimilitud es el mismo.

De igual forma los estadísticos que se obtiene de la tabla de confusión son levemente menores que el otro modelo logit, aunque siguen siendo buenos, así lo demuestra el Recall, es decir, el estadístico que clasifica correctamente a los postulantes que si se matricularon, los que predice el modelo, de aquellos que se matricularon pero el modelo predice que no. De igual forma los estadísticos F-Measure y Precision son menores que el modelo logit anterior.

A continuación se presenta la curva ROC para el modelo.

Gráfico 9: Curva ROC modelo.



Fuente: Elaboración propia.

La curva ROC del modelo es muy similar, prácticamente poseen la misma área bajo la curva (indicar porcentaje), levemente inferior para este modelo.

En la tabla 66 se presenta la clasificación de los matriculados por tramo de probabilidad para la base de entrenamiento.

Tabla 66: Clasificación matriculados y no matriculados base entrenamiento.

Probabilidad	Matriculados	No Matriculados
0%-10%	9%	91%
11%-20%	20%	80%
21%-30%	28%	72%
31%-40%	37%	63%
41%-50%	44%	56%
51%-60%	56%	44%
61%-70%	66%	34%
71%-80%	77%	23%
81%-90%	89%	11%
91%-100%	93%	7%

Fuente: Elaboración propia.

De la tabla se aprecia que este modelo clasifica con baja probabilidad de matrícula un 9% de alumnos que si se matriculan, un 1% más que el modelo que no cuenta con interacciones, de igual forma, asigna a una mayor cantidad de postulantes, 7%, una alta probabilidad de matrícula siendo que estos no se matriculan finalmente.

Analizando las características en común del grupo de postulantes que tienen una alta probabilidad de matrícula, pero que no se matriculan en la universidad son por ejemplo que el 97% de ellos postula en primera preferencia y el resto lo hace en segunda opción. Su puntaje pondera en general se encuentra sobre el promedio, y el 55% de ellos proviene de un colegio particular. Estas tres variables tienen una alta importancia en el modelo, son de las variables que más influyen en la probabilidad de matrícula. Todos ellos simularon beneficio en el portal web y el 83% de ellos lo obtuvo, y todos emitieron el certificado para hacer efectivo el beneficio. Además postulan a carreras que presentan una alta conversión anterior, que es la variable más influyente en el modelo. Todas estas características hacen que el modelo sobreestime la probabilidad de estas personas, que son buenos postulantes, pero finalmente no se matriculan.

En relación al grupo que tiene una baja probabilidad de matrícula las características que tienen en común estos postulantes es que ninguno postula en la primera preferencia a la universidad, y el 54% lo realiza entre la segunda y tercera, por lo que eso disminuye sus probabilidades de matrícula, la mayoría de ellos tiene un puntaje ponderado bajo el promedio y el 92% proviene de colegio subvencionado o municipal, por lo que en este ítem también el modelo le asigna baja probabilidad de matrícula. Junto a esto un 15% realiza simulaciones para obtener beneficios y todos ellos obtuvieron beneficios, pero ninguno emitió el certificado para validar el beneficio, además, solo un 8% lo realiza en la misma carrera a la que fue convocado finalmente. EL 69% pertenece al quintil de menores ingresos y postulan en su mayoría a carreras con baja conversión. El 80% postula a carreras que pertenecen al grupo del 20% con más baja conversión, por lo que esto afecta en mayor medida a que el modelo les asigne una baja probabilidad y no los clasifique en forma adecuada.

En la tabla 67 se puede apreciar la cantidad de postulantes en cada uno de los tramos de probabilidad.

Tabla 67: Conversión real y estimada por el modelo logit con interacciones.

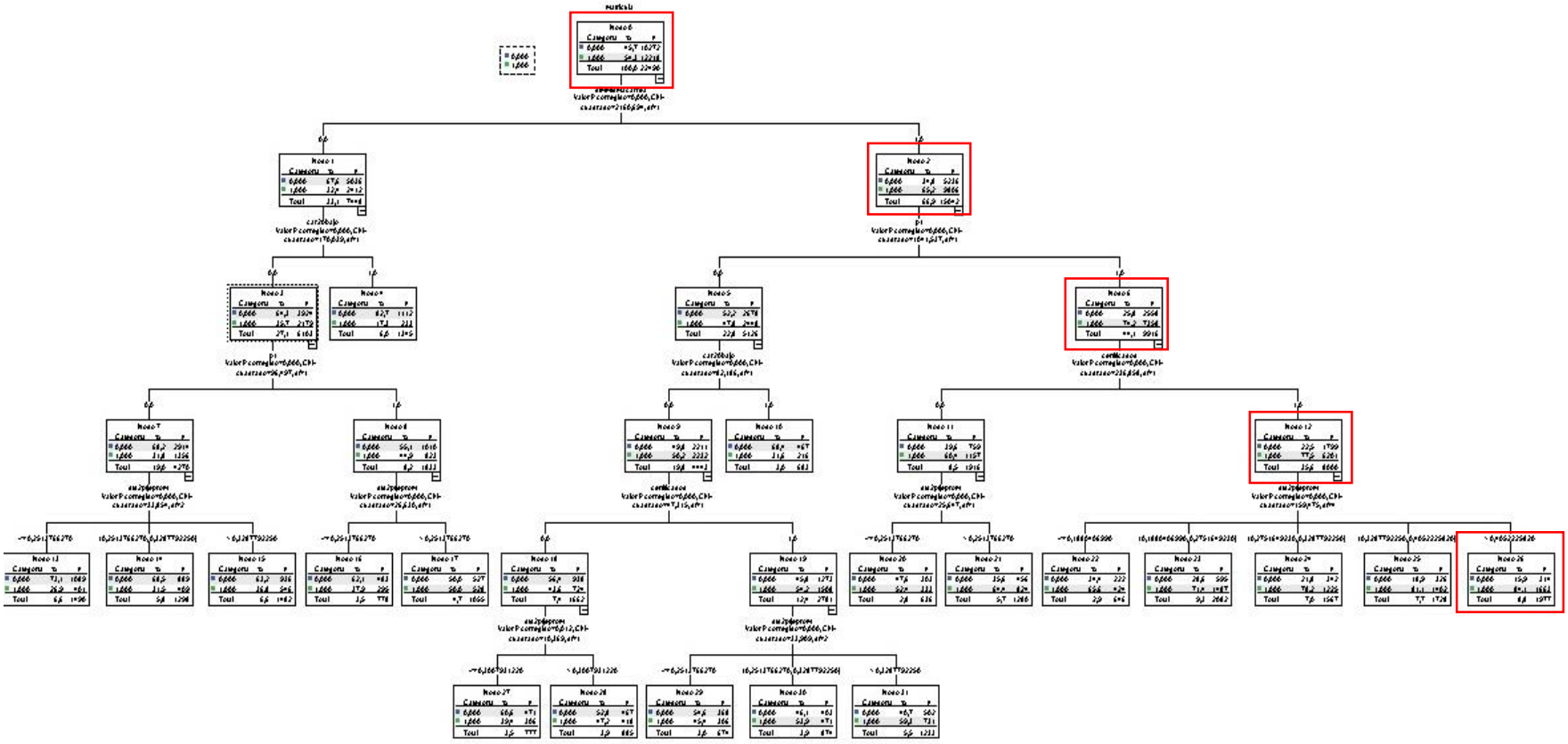
Tramo probabilidad	Conversión Real	Conversión estimada	Error	Cantidad de postulantes
0%-10%	11,9%	7,4%	4,50%	117
11%-20%	18,3%	15,5%	2,80%	1058
21%-30%	28,3%	25,2%	3,10%	1863
31%-40%	36,9%	34,9%	2,00%	2596
41%-50%	46,3%	44,9%	1,40%	3215
51%-60%	55,9%	54,9%	1,00%	3954
61%-70%	67,1%	65,2%	1,90%	5327
71%-80%	78,3%	75,1%	3,22%	6400
81%-90%	89,9%	84,4%	5,50%	6381
91%-100%	90,6%	91,8%	1,20%	1170

Fuente Elaboración propia.

9.11 ANEXO 11: Árbol de Decisión completo. Referenciado en punto 4.5.3.2 Resultados del modelo.

En rojo se encuentra marcado los nodos que lleva al nodo final, según el estadístico chi-cuadrado.

Ilustración 4: Árbol de Decisión.



Fuente: Elaboración propia.